CONSEQUENCES OF POWER TRANSFORMS IN LINEAR MIXED-EFFECTS MODELS OF CHRONOMETRIC DATA

Van Rynald T. Liceralde

A thesis submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Psychology and Neuroscience (Cognitive Psychology).

Chapel Hill
2018

Approved by:

Jennifer E. Arnold

Patrick J. Curran

Peter C. Gordon

ABSTRACT

Van Rynald T. Liceralde: Consequences of Power Transforms in Linear Mixed-Effects Models of Chronometric Data
(Under the direction of Peter C. Gordon)

Psycholinguists regularly use power transforms in linear mixed-effects models of chronometric data to avoid violating the statistical assumption of residual normality. Here, I extend previous demonstrations of the consequences of using power transforms to a wide range of sample sizes by analyzing word recognition megastudies and performing Monte Carlo simulations. Analyses of the megastudies revealed that stronger power transforms were associated with greater changes in fixed-effect test statistics and random-effect correlation patterns as they appear in the raw scale. The simulations reinforced these findings and revealed that models fit to transformed data tended to be more powerful in detecting main effects in smaller samples but were less powerful in detecting interactions as sample sizes increased. These results suggest that the decision to use a power transform should be motivated by hypotheses about how the predictors relate to chronometric data instead of the desire to meet the normality assumption.

To all my teachers, great and horrible.
To 17-year old Van who thought that he's not meant to learn to write code.
It was not you. You have it in you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANOVA	Analysis of variance
ELP	English Lexicon Project
FPP	Form Priming Project
Freq	word frequency
Len	word length
LMM	linear mixed-effects model
ms	milliseconds
Prime	form of prime; represents the form priming effect
QQ	quantile-quantile
Rel/Relate	word relatedness; represents the semantic priming effect
RT	response time
SPP	Semantic Priming Project
sqrt	square root
Voc/Vocab	vocabulary

LIST OF SYMBOLS

y	the dependent/outcome/criterion measure
y^*	the transformed dependent/outcome/criterion measure
λ	Box-Cox transformation parameter
x_p	the pth item-level predictor
z_q	the qth person-level predictor
$x_p x_{p'}$	same-level two-way interaction of two item-level predictors
$x_p z_q$	cross-level two-way interaction of an item-level predictor and a person-level predictor
$x_p x_{p'} z_q$	cross-level three-way interaction of two item-level predictors and a person-level predictor
γ	fixed-effect parameter
u	random effect (intercept or slope)
$\boldsymbol{T_{subj}}$	variance-covariance matrix of the random subject effects
τ_{00k}	random item intercept variance
ε_{ijk}	residual for trial *i* , made by subject *j* on item *k*
σ^2	trial-level residual variance

Introduction

Time is a hallmark measure used by cognitive scientists to make inferences about the architecture of the mind. Most models and methods examining basic cognition assign a central role to time, such that a prerequisite for any good model of cognition is to account for the time it takes to respond to a stimulus under various conditions. The ubiquity and importance of time as a measure of cognitive processing underlies the motivation to analyze chronometric data appropriately to make valid inferences and develop accurate models of cognition.

A satisfactory analysis of chronometric data requires appropriately accounting for its consistent positive skew (Luce, 1986). Given that standard linear statistical models assume that the residuals – and by extension, the outcome measure – are normally distributed in the population (i.e., *normality* assumption), analyses of chronometric data typically violate this assumption. When fitting linear mixed-effects models (LMMs), violating the normality assumption has been shown to produce unreliable standard errors and inference tests with potentially questionable results (Raudenbush & Bryk, 2002; Maas & Hox, 2004). Therefore, to preclude violating the normality assumption, it has been recommended from a statistical standpoint to apply a power transformation to chronometric data (e.g., the log- or inverse-transform; Box & Cox, 1964) in order to approximate normality and potentially increase the power of statistical tests (Judd, McClelland, & Culhane, 1995; Levine & Dunlap, 1982; Ratcliff, 1993).

However, some researchers have revived the argument (Sternberg, 1969; Loftus, 1978) that transforming time may produce unintended and unwanted theoretical and practical consequences (Balota, Aschenbrenner, & Yap, 2013; Lo & Andrews, 2015). Most notably, power transforms systematically alter units of duration, thereby re-expressing how RTs change as a function of the predictors in the model. Comparing models of raw and transformed data is difficult because effects in the raw scale would not necessarily (or at the very least, identically) manifest in the transformed scale. Because of this, questions

about cognitive processes, which occur over time, may not be answered directly when their scale of measurement is changed by a power transform (Lo & Andrews, 2015). This presents a dilemma: while we seek to meet the assumptions of linear modeling through power transforms, we also want our analyses to reveal information about cognitive processes, which occur in the raw scale. As Lo & Andrews (2015) aptly put it,

> "Cognitive psychologists are therefore trapped between a rock and a hard place. Analyses on raw RT [response times] are inappropriate because they fail to meet the assumptions of the linear model, but analyses on transformed RT are uninformative because they fail to answer the research questions of interest." (Lo & Andrews, 2015; p. 3)

The value of using power transforms to address violations of normality in LMMs is therefore challenged by how they potentially affect the inferences that are drawn about cognitive processes. This calls for a thorough assessment of the conceptual and statistical implications of power transforms, specifically as they are applied in LMMs of chronometric data. To this end, the current project evaluated the influence of power transforms on LMMs fit to RT data. I analyzed raw and transformed RT data from three word-recognition megastudies and I used parameters from the models I fit to simulate realistic datasets with different sample sizes and levels of variability. By analyzing several massive datasets and performing realistic simulations, converging evidence about the nature of power transforms' influence on LMMs fit to RT data can be obtained.

Linear Mixed-Effects Models

Analyses of chronometric data were traditionally performed on means (e.g., mean RTs, mean differences in RTs between conditions). Despite the consistent positive skew in chronometric distributions, violating the normality assumption in such analyses caused little concern because the sampling distribution of RT means converges on a normal distribution according to the Central Limit Theorem, implying that aggregate/mean RT analyses are robust to violations of normality (Glass, Peckham, & Sanders, 1972; Levine & Dunlap, 1982; Schmider, Ziegler, Danay, Beyer, & Bühner, 2015).

However, cognitive psychologists' studies typically involve repeated measures, where subjects respond to multiple items (e.g., words, sentences) and the same items are responded to by multiple subjects. Given that some individuals overall respond faster than others and some items are overall more quickly responded to than others, observations from repeated-measures designs – and the means computed from them – tend to be non-independent. Because standard linear modeling assumes that observations are independent, non-independence of observations from repeated-measures studies challenge the validity of mean RT analyses performed in those studies.

Non-independence of repeated observations has been widely recognized for random subject samples, but the same issue for item/stimulus samples has only been traditionally attended to in psycholinguistics. In particular, it has been routine for psycholinguists to address the potential issue that the effects they observe are not due to manipulations they introduce but instead due to specific idiosyncracies of the stimuli they selected for the experiment. Over the years, researchers have developed techniques, such as min *F*' (Clark, 1973) and $F_1 \times F_2$ ANOVAs (Raaijmakers, Schrijnemakers, & Gremmen, 1999), to address random variability from observations made on item samples. However, these aggregate techniques do not satisfactorily deal with dependencies introduced by repeated measures. Analyses of mean RTs collapse the variability in the RTs due to these dependencies and dismiss it as noise. By not accounting for these dependencies, aggregate analyses understate the variance present in the data because they separately consider between-subject and between-item variance while disregarding within-subject/-item variance, thereby inflating test statistics and increasing Type I error (Raudenbush & Bryk, 2002).

To address these dependencies, many psycholinguists have shifted in the last decade to using LMMs to analyze RT data (Baayen, Davidson, & Bates, 2007; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). LMMs extend the standard linear regression model by explicitly specifying nesting or cross-classification in the data through different levels of sampling, where observations are sampled from items and subjects, which are sampled from their respective populations. In doing so, LMMs distinguish item-related predictors (also called Level 1 predictors; x) from subject-related predictors (also called

Level 2 predictors; *z*), thereby allowing influences on RTs to be decomposed into between-item/within-subject and within-item/between-subject effects respectively.

LMMs address non-independence in data from repeated-measures RT studies by allowing the simultaneous estimation of random variability in RTs due to subjects and items. Subject and item samples potentially exhibit not only variability in their mean RTs (i.e., some subjects are overall faster than others/some items are overall responded to faster than others; also called *random intercepts*), but also variability in the strength with which they exhibit effects (i.e., the effect of some predictor is bigger for some subjects/some items than others; also called *random slopes*). Moreover, subjects' overall speed could be related to how strongly they exhibit certain effects (e.g., slower subjects tend to exhibit bigger word frequency effects (Yap, Balota, Sibley, & Ratcliff, 2011); also called *intercept-slope covariance*). Random intercepts, random slopes, and intercept-slope covariances – collectively called *random effects* – could all be specified and estimated in LMMs, and Table 1 shows how LMMs extend the standard linear regression model to account for all these sources of dependence and variability.

Model	Equations		Assumptions
Standard Linear Model (includes ANOVAs)	$y_{ijk} = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 z_{1j} + \beta_4 x_{1k} x_{2k} + \beta_4 x_{1k} z_{1j} + \varepsilon_{ijk}$	(main effects) (interaction effects) (residuals)	1) ε_{ijk} is independent and identically normally distributed (i.e., equal variance) with mean 0 and variance σ^2 ($\varepsilon_{ijk} \sim N(0, \sigma^2)$). But this assumption is violated because ε_{ijk} that come from the same subject *j* or made on the same item *k* are going to be correlated. 2) The magnitudes of predictors' effects are identical across items and across subjects
Linear Mixed-Effects Model (LMM)	Level 1: $y_{ijk} = \beta_{0jk} + \beta_{1j} x_{1k} + \beta_{2j} x_{2k} + \beta_3 x_{1k} x_{2k} + \varepsilon_{ijk}$ Level 2: $\beta_{0jk} = \gamma_{00} + \gamma_{01} z_{1j} + u_{0j} + u_{0k}$ $\beta_{1j} = \gamma_{10} + \gamma_{11} z_{1j} + u_{1j}$ $\beta_{2j} = \gamma_{20} + u_{2j}$ $\beta_3 = \gamma_{30}$ Reduced form: $y_{ijk} = \gamma_{00} + \gamma_{10} x_{1k} + \gamma_{20} x_{2k} + \gamma_{00} z_{1j} + \gamma_{11} x_{1k} z_{1j} + \gamma_{30} x_{1k} x_{2k} + u_{0j} + u_{0k} + u_{1j} x_{1k} + u_{2j} x_{2j} + \varepsilon_{ijk}$ Matrix form: $\boldsymbol{y} = \boldsymbol{X\gamma} + \boldsymbol{Su_{subj}} + \boldsymbol{Wu_{item}} + \boldsymbol{\varepsilon}$	(main effects) (interaction effects) (random intercepts) (random slopes) (residuals)	1) Means may vary across Level 2 units (random intercepts) • $u_{0j} \sim N(0, \tau_{00j})$ and $u_{0k} \sim N(0, \tau_{00k})$ 2) The magnitude of predictors' effects may vary across Level 2 units (random slopes) • $u_{ij} \sim N(0, \tau_{11j})$ and $u_{2j} \sim N(0, \tau_{22j})$ 3) Variability in the means may be related to the variability in the magnitude of predictors' effects (intercept-slope covariance) • $\tau_{01j} = correlation(u_{0j}, u_{ij}) \times \left(\sqrt{\tau_{00j}}\right)\left(\sqrt{\tau_{11j}}\right)$ 4) Collectively, these random effects are independent and identically have a multivariate normal distribution with mean 0 and variance-covariance matrix $\boldsymbol{T}$ • $\boldsymbol{U_{subj}} \sim MVN\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}\tau_{00j} & & \\ \tau_{01j} & \tau_{11j} & \\ \tau_{02j} & \tau_{12j} & \tau_{22j}\end{bmatrix}\right)$ 5) Variability between observations due to nesting or cross-classification (e.g., $\boldsymbol{T_{subj}}$ and τ_{00k}) is independent of residual variability (i.e., $\varepsilon_{ijk} \sim N(0, \sigma^2)$)

Table 1. Overview of LMMs extending the standard linear model. Note. y_{ijk} is the response to trial i by subject j on item/word/sentence k; x refers to the Level 1 predictors; z refers to the Level 2 predictors; γ refers to the fixed-effect coefficients in the LMM; u refers to the random effects; and ε_{ijk} refer to the Level 1 residuals. For the matrix form, $\boldsymbol{X\gamma}$ refers to fixed-effects component of the model and $\boldsymbol{Su_{subj}} + \boldsymbol{Wu_{item}}$ refer to the random-effects component of the model.

Unlike aggregate analyses, LMMs are not necessarily robust to violations of the normality assumption. Little is known about the severity of the consequences of having non-normal Level 1 residuals (ε_{ijk}), but this non-normality is thought to affect the computation of standard errors for fixed effects (i.e., γ) at both levels of the model (Raudenbush & Bryk, 2002), and having a big sample size does not necessarily protect against this influence (Goldstein, 2011). On the other hand, having non-normal Level 2 residuals (u_{0j}, u_{0k}) typically only produces highly inaccurate standard errors of Level 2 random effects (e.g., τ_{00j}, τ_{00k}; Maas & Hox, 2004), but if the non-normal distribution has outliers or heavy tails, its effects also seep into the standard errors of Level 2 fixed effects (Raudenbush & Bryk, 2002; Seltzer, 1993). Coupled with few Level 2 units (e.g., subjects and/or items), non-normal Level 2 residuals could also influence confidence intervals for fixed effects at both levels (Maas & Hox, 2004). Suffice it to say that violations of normality (and other standard linear modeling assumptions) may compromise the validity of inferential tests in unpredictable and disconcerting ways.

Box-Cox Transforms

To avoid violating the normality assumption when fitting LMMs to RTs, psycholinguists in some instances apply the log or inverse transform to the RTs prior to data analysis to approximate normality (e.g., Baayen, 2008; Baayen & Milin, 2010; Masson & Kliegl, 2013). The log and inverse transforms are special cases of the Box-Cox family of power transforms, which is generally expressed via the following piecewise function:

$$y^* = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln(y), & \lambda = 0 \end{cases}$$

where the power parameter λ that would optimally normalize the data is estimated using a profile likelihood function (Box & Cox, 1964). Table 2 summarizes some of the power transforms commonly used to normalize RTs and the values of λ for the Box-Cox function that would roughly correspond to these transforms. Note that because skew varies across datasets, the power transform needed to optimally

normalize the data also varies across datasets. In the context of the Box-Cox procedure, the farther the estimated λ is from 1, the stronger the power transform required to optimally normalize the data. Thus, datasets optimally normalized by the inverse transform according to the Box-Cox procedure will not be sufficiently normalized by the log transform, whereas datasets optimally normalized by the log transform will be "over-transformed" by the inverse transform so as to restore, if not worsen, the skew of the dataset.

Power Transform	Equation	Box-Cox Parameter (λ)
Identity	$y^* = y$	1
Square-root	$y^* = \sqrt{y}$	0.5
Log	$y^* = \ln(y)$	0
Inverse Square-root	$y^* = \frac{-1000}{\sqrt{y}}$	-0.5
Inverse	$y^* = \frac{-1000}{y}$	-1

Table 2. Commonly used power transforms to normalize RTs, along with their corresponding equations and the Box-Cox parameters that would approximate these power transforms. A value of -1000 is multiplied to the inverse square-root and inverse transforms to magnify the transformed values and to maintain the rank-order of the RTs.

While it is straightforward to use a Box-Cox transform to optimally normalize RTs, the conceptual and statistical implications stemming from its use are much less so. In fact, rashly transforming RTs could lead researchers to obtain uninterpretable results and make invalid conclusions about cognition at the cost of carelessly meeting the normality assumption (Lo & Andrews, 2015).

Implications of Using Power Transforms

Chronometric data have been successful measures of cognitive processing because they stand as one of the most direct measures of unobservable mental processes. Since cognitive processes occur over time, it is reasonable to consider that the properties of time as a ratio measure of physical events may generalize to the measurement of mental events. That said, chronometric data may at least have interval properties when used to measure cognitive processing. In fact, the appeal of methods that use time to characterize stages of information processing – such as the subtraction method (Donders, 1969) and the

additive factors method (Sternberg, 1969) – is based on chronometric data's equal-intervals property. The logic of these methods is that the change in RT due to manipulations added in an experiment indicates the necessary processing stages the mind undergoes to respond to those manipulations and, consequently, the amount of time it takes to engage in those stages. In this framework, changes in RT reveal something about the sequence and independence of processing stages the mind undertakes when presented some stimulus. Two stages are considered independent in their output if the effects of manipulations tapping into those stages do not modulate each other (i.e., they do not interact) as observed in the RT data. Researchers have inferred the independence or interaction of processing stages from RTs because its equal-intervals property allows the direct comparison of RTs from different experimental conditions.

However, power transforms distort RTs' equal consecutive intervals. The most notable consequence of this distortion is that by rescaling RTs, differences in RTs across experimental conditions are no longer easily comparable. This is because consecutive intervals in the log or inverse scale are not equal in the way that consecutive intervals in raw/untransformed RTs are. In the raw scale, differences in short and in long RTs are weighted equally, allowing for direct comparison of these differences. In contrast, by expanding the short end of the RT distribution and compressing the long end, power transforms exaggerate differences in short RTs and downplay those in long RTs. Thus, log/inverse-scale differences observed in one part of an RT distribution cannot be directly compared to those observed in another part: interpreting the magnitude of a difference in the log or inverse scale depends on where this difference is located in the RT distribution. For example, a 0.2877 log-difference could be a 100-ms difference in the short end of an RT distribution or a 300-ms difference in the long end. Alternatively, an identical 200-ms raw difference in the short and long ends of the same RT distribution results in distinct differences in the log scale (Table 3).

The RT differences are…	Part of RT Distribution	Raw	Log
Different in raw, identical in log	Short	$400 - 300 = 100\,ms$	$\ln(400) - \ln(300) \sim 0.2877$
	Long	$1200 - 900 = 300\,ms$	$\ln(1200) - \ln(900) \sim 0.2877$
Identical in raw, different in log	Short	$500 - 300 = 200\,ms$	$\ln(500) - \ln(300) \sim 0.5108$
	Long	$1200 - 1000 = 200\,ms$	$\ln(1200) - \ln(1000) \sim 0.1823$

Table 3. Illustration of how power transforms can obscure differences in the raw RT scale and introduce differences absent in the raw RT scale.

The distortion of raw RTs' interval properties by power transforms leads to theoretical challenges. For one, the distortion affects researchers' ability to characterize the independence or interaction of stages of information processing. Identical RT differences between manipulations in the raw scale (which would suggest the independence of processing stages) could present as unequal RT differences in the transformed scale (which would suggest an interaction between processing stages) (Loftus, 1978). Thus, results obtained from transformed data could lead to completely different conclusions from results obtained from raw data about whether the processing stages tapped by manipulations in an experiment are independent or interacting.

To illustrate, in the word recognition literature, word frequency and stimulus quality have been argued to have additive/independent effects on word recognition based on analyses of raw RTs (Borowsky & Besner, 1993; Yap & Balota, 2007; Yap et al., 2008), which show equivalent increases in RTs to visually degraded words, as compared to visually intact words, for high- and low-frequency words. However, mean RT analyses disregard variability in RTs due to individuals dynamically adjusting their processing across an experiment (Kinoshita, Forster, & Mozer, 2008; Masson & Kliegl, 2013). Masson & Kliegl (2013) argued that given RTs' sensitivity to properties of previous trials, the interaction of word frequency and stimulus quality might manifest across trials, and mean RT analyses might be obscuring this interaction. Indeed, Masson & Kliegl found this interaction when they incorporated the previous trial's word frequency and stimulus quality into LMMs fit to RTs that had been inverse transformed to approximate residual normality. However, Balota et al. (2013) found that fitting LMMs to inverse-transformed RTs produced significant interactions between word frequency and stimulus quality

across trials in two of three published datasets when those interactions were not obtained in the original ANOVAs nor in LMMs fit to raw RTs. While analyses on raw RTs suggest that the processing stages tapped by word frequency and stimulus quality are additive/independent (e.g., 650 ms – 550 ms = 100 ms for high frequency words; 850 ms – 750 ms = 100 ms for low-frequency words), inverse-transformed RTs could suggest otherwise (e.g., (-1000/650) – (-1000/550) = 0.28 for high frequency words; (-1000/850) – (-1000/750) = 0.16 for low frequency words). With a sufficiently large sample, this interaction could manifest as significant when in fact it does not exist in the raw RT scale in the population.

These mathematical and empirical examples show how the distortion introduced by power transforms could lead to different conclusions about the independence and interaction of processing stages. In fact, Sternberg (1969) himself claimed that "additivity will in general be destroyed by nonlinear transformation of measurements" (p. 286).

Another theoretical challenge brought about by the scale distortions introduced by power transforms concerns distributional analyses of chronometric data. Researchers posit that different cognitive processes are associated with different components of the chronometric distribution (Balota & Yap, 2011). Consequently, factors that influence these cognitive processes should show their effects in the respective components of the distribution. For instance, the skewed tail of the chronometric distribution has been associated with decision processes, where tasks that require more difficult decisions generate chronometric distributions with greater skew (Ratcliff & Murdock, 1976). In turn, age, IQ, working memory, and other individual difference factors, which have been demonstrated to influence the speed of the decision process (Ratcliff, Thapar, & McKoon, 2006; 2010), have been shown to affect the long/slow tail of the distribution (Schmiedek et al., 2007; Tse et al., 2010). On the other hand, the modal part of the distribution has been associated with rather early and automatic processes, reflecting some cognitive/neural/motor preactivation that may or may not be stimulus-driven (e.g., Balota & Spieler, 1999; Yap, Balota, & Tan, 2013). In turn, word frequency, semantic priming, and other lexical and stimulus factors, which are argued to influence early recognition processes (Brown & Hagoort, 1993; Sereno, Rayner, & Posner, 1998) have been shown to influence both the modal part of the distribution

(and the slow tail of the distribution, depending on the task; Hoedemaker & Gordon, 2014; 2017; Staub et al., 2010; Yap, Tse, & Balota, 2009).

Given that individual differences and lexical factors exhibit their effects in distinct components of the RT distribution, power transforms can influence the ability to detect and estimate the effects of both types of factors. By expanding the short end and compressing the long end of the RT distribution, power transforms can prevent researchers from finding effects of individual-difference factors which primarily affect the long/slow end, while exaggerating the effects of lexical factors which primarily affect the short/fast end. Thus, power transforms could downplay the role of individual differences in the data while overstating the magnitude of the effects of lexical factors.

However, dismissing the utility of power transforms based solely on these theoretical challenges is premature. The success of a statistical model is ultimately determined by how accurately it accounts for patterns in the population, and perhaps implicit in the statistical recommendation to apply a power transform to RTs is the assumption that the LMM fit to transformed data is more likely to achieve this correspondence than the raw RT LMM. Researchers seem to think that the mechanism by which power transforms facilitate this correspondence is by meeting statistical assumptions: in the literature, researchers tend to justify transforming RTs simply to make the RT distribution normal. However, an underappreciated mechanism by which transformed models might reflect the population more closely is that they characterize the relationship between predictors and RTs with a different functional form. That is, while cognitive processes occur in the raw chronometric scale, transformed LMMs indicate that the predictors do not influence the time it takes for a process to occur *per se* but rather some function of time.

For example, comparing the magnitude of absolute differences in RTs between younger and older adults is not necessarily valid because older adults' RTs tend to be confounded by the general cognitive slowing associated with aging (Salthouse, 1985). By log-transforming the RTs, this confound is reduced and the effects of the predictors are instead characterized as a proportional change (i.e., log(900) – log(800) = log(900/800)) in performance instead of an absolute difference (900 – 800) (Lo & Andrews, 2015). A researcher might also hypothesize that some manipulation affects the general rate of a cognitive

process and not just the duration of individual events that engage this process, in which case fitting an LMM to inverse-transformed RTs might be appropriate to model this relationship. If the underlying relationships of the predictors with RTs take on these different functional forms in the population, the LMMs fit to transformed RTs might be better suited to characterize these relationships than those fit to RTs.

Collectively, issues surrounding the use of power transforms have prompted some researchers to suggest analyzing both the raw and transformed data instead of just analyzing one type of RT (Wagenmakers, Krypotos, Criss, & Iverson, 2012). If effects are consistent across both types of data, this is taken as converging evidence that the effect is stable, and analyses on the raw data are reported for ease of interpretation. Another approach researchers take if effects are significant on the transformed data is that they back-transform the data to the raw scale and then create figures which show the "significant" effect on the raw scale. However, these approaches do not avoid the issues generated by power transforms because significant effects in the transformed scale do not guarantee that the same effects are significant in the raw scale (Lo & Andrews, 2015; Berry, DeMeritt, & Esarey, 2010).

Reporting results for both types of RTs does not reveal which model better represents how the effects manifest in the population, particularly in cases where LMMs fitted to raw and transformed RTs reveal inconsistent results. On the one hand, the raw RT model might be preferable because it tends to be the conceptually appropriate model given that cognitive processes occur in the raw scale. However, inferential tests based on this model are questionable because it violates the normality assumption. In contrast, the transformed model might be preferable because by meeting statistical assumptions, the model generates parameter estimates for whom inference tests are statistically valid. But how useful are statistically valid inference tests if the inferences made do not correspond to the phenomena about which they are made?

The tradeoff between conceptual and statistical appropriateness in analyzing raw vs. transformed RTs introduces a dilemma about the use of power transforms. This dilemma is particularly problematic when LMMs fit to raw and transformed RTs produce divergent results because these results could lead to

mutually exclusive inferences about phenomena in the population. This underscores the motivation to assess the extent to which the raw and transformed LMMs produce divergent results. Specifically, given that the success of a statistical model is determined by how accurately it reflects and predicts phenomena in the population, it is important to identify which between the raw and transformed LMM more likely and more closely corresponds to the population when they reveal inconsistent results. I investigated these issues through simulations, where I specified the underlying functional form of the predictors' relationship with RTs and assess the influence of power transforms on model results.

Setting Up the Simulations

Simulation is a well-established method in psychological research. Simulations are economical and efficient tools for assessing the long-run performance of various statistical techniques and for modeling phenomena without having to invest resources in collecting experimental data. However, simulations are only useful and successful insofar as they are informative; they can fail and therefore offer little or no scientific value if they insufficiently correspond to relevant aspects of the reality they are attempting to imitate (Grim et al., 2011).

Successful simulations of RT data need to correspond to several critically relevant features of observed RT data. First, RTs are not arbitrary and are taken to indicate the amount of time it takes to engage a series of cognitive processes for some behavior to be observed. Second, human behavioral data exhibits a substantial amount of variability and covariance. Third, psycholinguists tend to fit rather complex models to their data, incorporating a substantial number of language-related and individual-difference predictors, while sometimes keeping to the statistical recommendation to maximize the random-effect structure of the LMM as warranted by the study design (Barr, Levy, Scheepers, & Tily, 2013).

Therefore, to be consistent with observed RT data, the simulations would need to consist of meaningful parameters, to be fairly complex in their data-generating mechanisms, and ultimately to generate data that resemble observed RTs. That is, the parameters used in the simulation should be based

on reasonable estimates of the magnitude of predictors' effects and of variability present in typical (word-recognition) RT data. Moreover, the data-generating process should reflect some of the complexity that psycholinguists believe to exist in their data, as implied by the complex LMMs they fit to their data. For instance, including several random slopes for predictors in an LMM implies not only that there is variability in the magnitude of predictors' effects across subjects/items, but also that these predictors' effects potentially covary. A meaningful simulation of RT data should then exhibit some of these complexities and produce RTs that are a composite of different sources of variability.

To meet these criteria, I analyzed data from several megastudies and used the estimates obtained from these analyses to set up the simulations. Megastudies are massive, multisite efforts to develop comprehensive behavioral databases for broad sets of linguistic stimuli (see Keeulers & Balota, 2015 for an overview). Researchers have traditionally used them to randomly select stimuli that fit certain linguistic and behavioral profiles for use in experiments. Consisting of hundreds of thousands to millions of observations, megastudies have rich information about behavioral performance on linguistic stimuli (e.g., RT, accuracy), properties of linguistic stimuli (e.g., word frequency, word length, orthographic neighbors, prime type) and demographic and individual-difference characteristics for all subjects in the study (e.g., vocabulary, attention).

By analyzing megastudies, I can empirically demonstrate the influence of power transforms on LMMs fit to massive datasets, thus extending previous demonstrations in small datasets (Balota et al., 2013; Lo & Andrews, 2015). Moreover, the size and richness of megastudies make it possible to base the simulations on them. Their size guarantees that estimates obtained from analyzing them would make for meaningful and realistic parameters to be used in the simulations because these estimates would better approximate the magnitude of effects and variability in the population than estimates from small studies or arbitrary values. Their richness allows reasonably complex LMMs to be fit to the data, and these complex LMMs could be used to simulate RTs that are a composite of various meaningful sources of variability. By basing the simulations on megastudies, I expect the results obtained from the simulations to be reliable due to achieving sufficient correspondence with observed RT data.

Current Project

In the current project, I analyzed data from several megastudies and performed Monte Carlo simulations to evaluate the extent of power transforms' influence in LMMs fitted to RT data. The project consisted of three studies, and each study was based on trial-level data from a word-recognition megastudy. Study 1 was based on the *Semantic Priming Project* (SPP; Hutchison et al., 2013); Study 2 was based on the *Form Priming Project* (FPP; Adelman et al., 2014); and Study 3 was based on the *English Lexicon Project* (ELP; Balota et al., 2007).

Each study consisted of three parts (see Figure 1 for a roadmap):

1. ***Analysis of megastudy data***. Data from each megastudy were transformed using a log- ($\ln(RT)$), an inverse-square-root- ($-1000/\sqrt{RT}$), and an inverse-transform ($-1000/RT$), which in addition to the raw RT data resulted in four RT scales[1]. The data were then preprocessed for outliers and LMMs were fit to the remaining data, one for each RT scale. Afterwards, I used the Box-Cox procedure to estimate the λ that would optimally normalize the trial-level (i.e., cluster-specific) residuals, and then created residual QQ plots for each of the LMMs to visually inspect that the power transform which produced the least skewed residuals corresponded to the λ identified by the Box-Cox procedure. Finally, I compared the fixed- and random-effects results across the raw and transformed LMMs to determine the implications of reducing skew using power transforms.

   However, differences between the models do not indicate which model is "better" because the data-generating process in the population is unknown. Therefore, to assess the long-run performance of the raw and transformed models, I performed simulations based on the LMMs fit to the megastudy data and compared the models in how effectively they estimated the input conditions that were specified in the simulations.

[1]I also tested a shift transform ($RT - 200$) as a control transform. All results for this RT type were identical to the raw model results.

2. ***Simulations based on raw model as generating process*.** Small-version datasets of the megastudy were generated based on the raw RT LMM in Part 1 and the subject and item sample sizes within those simulated datasets were varied. Afterwards, the same raw and transformed models in Part 1 were fit to all the simulated data and then the performance of the models was assessed using several measures such as bias, coverage, convergence rates, and Type I error rate/power.

   The raw RT LMM was used as the basis for the generating process of the simulations because the majority of cognitive theories describe processing as occurring over time and in doing so, they assume (implicitly or explicitly) that the relationship between observed RT and the latent cognitive factors measured by the predictors is reasonably linear (Lo & Andrews, 2015; but see Wagenmakers et al., 2012).

3. ***Simulations based on optimally transformed model as generating process***. Another set of simulations was performed based on the LMM fit to the optimally normalized megastudy RTs, and the same performance and comparison measures in Part 2 were assessed. The purpose of these simulations is two-fold: first, they address the possibility that outcomes in Part 2 depended on using the raw LMM to simulate data. Second, they test whether there are advantages to fitting LMMs in the true scale of the generating process (e.g., would inverse LMMs perform better than raw LMMs when the inverse scale is the true scale of the functional relationship between the predictors and RTs?).

I focused on three predictors of RT and their interactions in each study. Two of these predictors were word-/item-level predictors (x_1 and x_2), whereas one was a person-level/individual difference predictor (z_1). I focused on these predictors because LMMs of psycholinguistic data typically involve controlling for item characteristics and individual differences to see whether manipulations of interest produced differences in behavior. Moreover, by using predictors at different levels, I examined the extent to which transforming RTs influenced various levels of effects (i.e., main effects, same-level [$x_1 x_2$] vs.

cross-level interactions [$x_1 z_1$]). Focusing on three predictors at different levels allowed complex LMMs to be fit to the simulated data while preventing non-convergence rates from increasing due to the maximal random-effect structure exponentially expanding as the number of predictors increase.

If results between the analyses of the megastudies and the Monte Carlo simulations are consistent, this would indicate that the influence of power transforms on main (additive) and/or interaction effects generalizes across subject and item sample sizes. More importantly, if the patterns of results across the three studies are consistent, it can be inferred that the influence of power transforms extend to RT data from different kinds of studies.

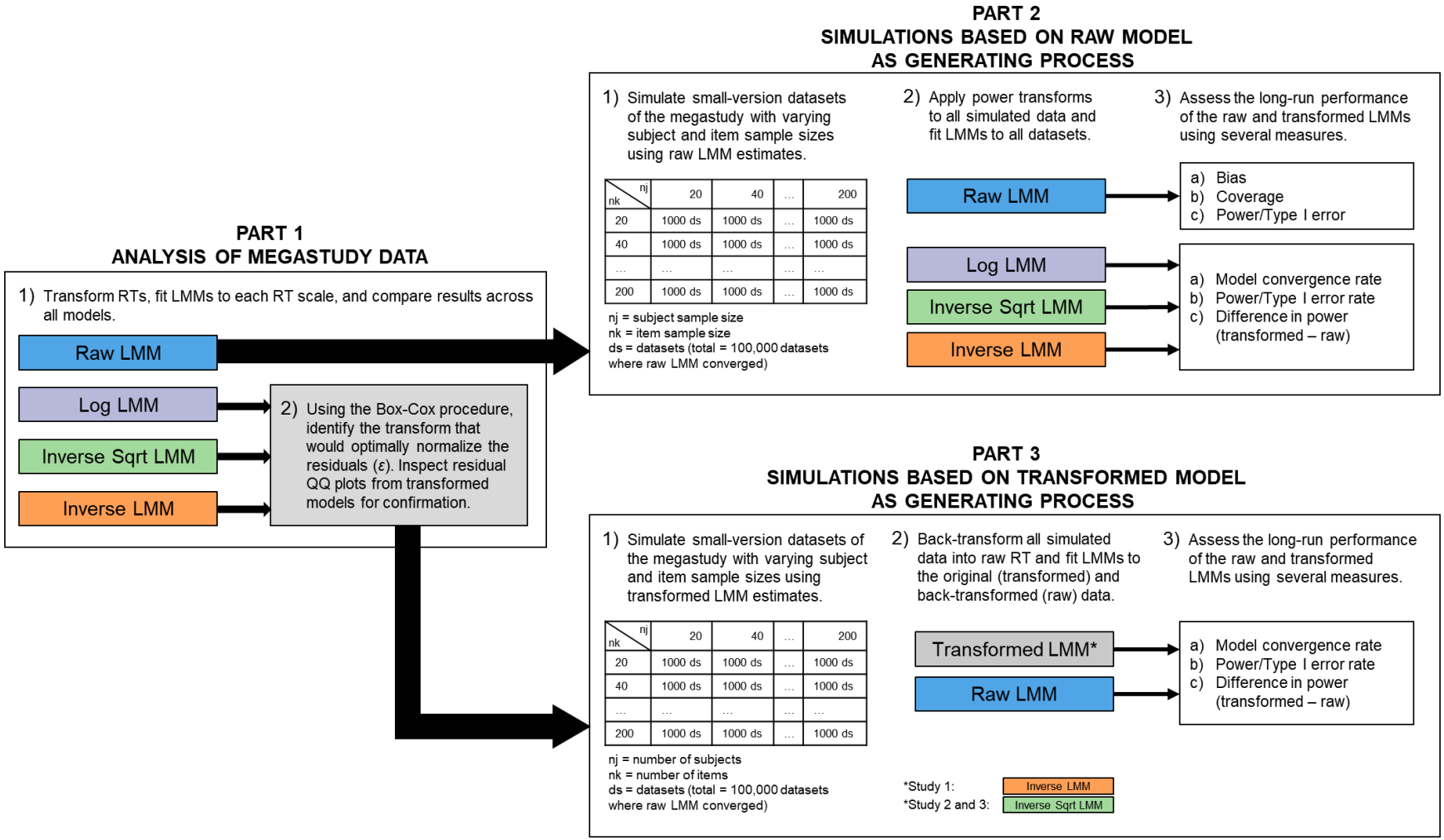

Figure 1. Roadmap of procedures for each of the three studies conducted in the current project.

Method

Study 1: Semantic Priming Project (SPP)

The SPP (Hutchison et al., 2013) is a megastudy of semantic priming in word recognition. In a typical semantic priming experiment, a target word (e.g., doctor) is briefly preceded (~50 ms) by a semantically related (e.g., nurse) or unrelated (e.g., bread) prime, and responses to target words are faster on average when preceded by a related prime (Meyer & Schvaneveldt, 1971). I analyzed the lexical decision subset of the SPP which consists of data from 512 subjects, each of whom responded to 830 or 831 prime-target word pairs and filled out a battery of individual-difference measures. More details about the procedure are available in Hutchison et al. (2013).

In Study 1, I focused on the following three predictors in the SPP and their interactions:

1) The target word frequency (x_1), which is a continuous word-level characteristic representing the target's $\log_{10}$ frequency per million words as obtained from SUBTLEX-US, a corpus of frequencies from subtitles of movies and TV shows (Brysbaert & New, 2009);
2) Prime relatedness (x_2), which is a word-level manipulation indicating whether the target was preceded by a semantically related or unrelated prime, thus representing the semantic priming effect; and
3) Vocabulary (z_1), which is a continuous person-level characteristic representing the average score of a subject out of three vocabulary subtests from the Woodcock-Johnson reading battery (Woodcock, McGrew, & Mather, 2001)

Study 1.1: Analyzing SPP data

Data preprocessing. Only RTs from accurate responses to lexical targets (i.e., word strings correctly identified as words) were analyzed. Prior to screening the data for outliers, word frequency and

vocabulary were item-mean and grand-mean z-transformed respectively[2]. The data were then preprocessed for outliers as follows: first, following Yap, Balota, Sibley, & Ratcliff (2012), observations that were shorter than 200 ms or longer than 3000 ms were excluded. Second, I applied the three power transforms (i.e., log, inverse square-root, and inverse) to the remaining data and then performed the non-iterative moving outlier criterion procedure recommended by Van Selst & Jolicoeur (1994) for each RT scale. For every subject, this procedure adjusts the outlier criterion by the sample size observed for each experimental condition to account for the uncertainty introduced by small sample sizes. The maximum bounds for this criterion are +/- 2.5 SDs from the mean for conditions that have sample sizes greater than 100 observations, and the criterion decreases with decreases in sample size. Third, subjects' mean RTs for each scale were recalculated and subjects whose updated mean RTs are shorter than 400 ms or longer than 1000 ms (and their transformed-RT equivalents) were considered unusual and were dropped from further analyses. Because each RT scale produced different means and standard deviations, the percentage of observations dropped from further analyses due to the outlier criteria described above range from 2.82% to 5.17% of the total data across RT scales. Five subjects and 15 target words were further excluded due to missing predictor values (word frequency, vocabulary), resulting in 1,646 items and 494 to 499 subjects being retained for analyses.

Fitting the LMMs and identifying the optimal power transform. The main effects of word frequency, semantic relatedness, and vocabulary and all interaction effects were entered into cross-classified LMMs, one for each RT scale. Random intercepts were specified for subjects and items and random subject slopes were specified for the word frequency (x_1) and relatedness (x_2) effects. All LMMs

[2]Each unique item contributed one data point in the calculation of the grand means and z-scores for word frequency and length; each subject contributed one data point in the calculation of the grand mean and z-scores for vocabulary. The z-scores for continuous predictors in the subsequent studies were also calculated this way. All test statistics obtained from fitting the LMMs using z-transformed predictors were identical to those that were obtained using mean-centered predictors.

(including those in subsequent studies) were fit with restricted maximum likelihood using the 1.1-12 version of the `lme4` package in R (Bates, Mächler, Bolker, & Walker, 2015; R Core Team, 2017)[3].

After the models were fit, the optimal power transform for the preprocessed data was identified using the `boxcox` function in the 7.3-45 version of the `MASS` package in R (Venables & Ripley, 2002; R Core Team, 2017). The Box-Cox procedure revealed that the preprocessed data is optimally normalized using λ = -0.95 which roughly corresponds to the inverse transform. Figure 2 shows that the residual QQ plots from the models supported the results of the Box-Cox procedure.

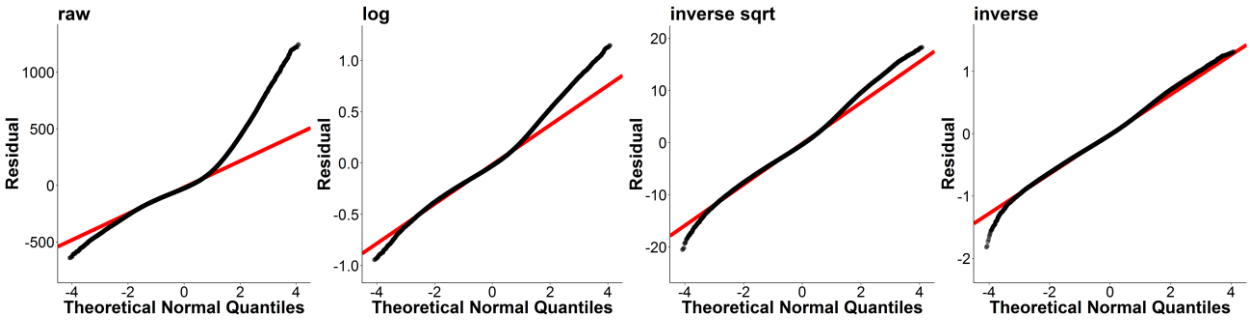

Figure 2. Trial-level residual QQ plots from models fit to SPP data

Because the four RT scales consist of different units of measurement, t-statistics were used as a standardized metric for comparing the models' fixed-effect estimates. Table 4 shows the t-statistics for each fixed effect across all four models. For the main effects, the t-statistics from the transformed models were larger than those from the raw model. However, for the two-way interactions, stronger power transforms resulted in greater changes to the t-statistics relative to the t-statistics in the raw model. Notably, the x_2z_1 interaction effect, whose t-statistic was distant from the critical value of $|2|$ in the raw scale, just surpassed this critical value in the inverse scale[4]. Moreover, although still significant, the t-

[3]The general lmer formula for all the LMMs run in this project is (`RT ~ x1*x2*z1 + (1 + x1 + x2|Subject) + (1|Item)`).

[4]Because calculating degrees-of-freedom and *p* values in cross-classified LMMs is non-trivial, I followed the convention of setting the critical *t* value $= |2|$ as a reasonable approximation of significance given $\alpha = 0.05$ and

statistics of the other two-way interactions moved closer to the critical value as more powerful transforms were applied to the data. Lastly, transforming the data did not affect the three-way interaction's t-statistic.

RT Scale	x_1 (Freq)	x_2 (Relate)	z_1 (Voc)	x_1x_2 (Freq x Relate)	x_1z_1 (Freq x Voc)	x_2z_1 (Relate x Voc)	$x_1x_2z_1$ (F x R x V)
Raw	-31.95	-28.65	-4.12	13.43	8.91	-0.48	0.04
Log	-33.85	-32.31	-4.18	13.48	8.61	-1.28	0.43
Inverse Sqrt	-34.04	-32.89	-4.39	13.11	7.55	-1.76	0.84
Inverse	-33.60	-32.20	-4.40	12.24	6.71	**-2.31**	0.54

Table 4. *t*-statistics of fixed effects from LMMs fit to the SPP data. Effects that changed significance from raw RT model are boldfaced. *Freq* = word frequency; *Relate* = semantic relatedness/priming effect; *Voc* = vocabulary

Beyond the fixed effects, the power transforms also affected the estimation of random-effect covariance patterns observed in the raw scale. Whereas larger word frequency effects were associated with slower subjects in the raw scale ($r = -0.48$), the opposite relation appeared in the inverse scale ($r = 0.36$). Moreover, while the magnitude of the semantic priming effect was not related to subjects' mean RTs in the raw scale ($r = 0.06$), the effect was shown to be smaller for slower subjects in the inverse scale ($r = 0.48$) (Table 5).

given that the t-distribution turns into the standard normal distribution as sample sizes increase (see Baayen (2008) as well).

	Raw LMM			Inverse LMM		
Fixed Effects	Estimate ($\hat{\gamma}$)	Std. Error	*t*	Estimate ($\hat{\gamma}$)	Std. Error	*t*
Intercept	664.01	5.01	132.47	-1.5975	0.0107	-149.09
x_1 (Frequency)	-29.73	0.93	-31.95	-0.0707	0.0021	-33.60
x_2 (Relatedness)	-21.26	0.74	-28.65	-0.0560	0.0017	-32.20
z_1 (Vocabulary)	-20.31	4.93	-4.12	-0.0464	0.0105	-4.40
x_1x_2 (Freq x Rel)	7.23	0.54	13.43	0.0132	0.0011	12.24
x_1z_1 (Freq x Voc)	5.00	0.56	8.91	0.0077	0.0011	6.71
x_2z_1 (Rel x Voc)	-0.35	0.74	-0.48	-0.0040	0.0017	-2.31
$x_1x_2z_1$ (F x R x V)	0.02	0.54	0.04	0.0006	0.0011	0.54
Random Effects	SD		Correlations	SD		Correlations
Subject ($\hat{\boldsymbol{T}}_{subj}$)						
Intercept	109.86	1		0.2355	1	
x_1	9.16	-0.48	1	0.0190	0.36	1
x_2	11.50	0.06	0.14	0.0307	0.48	0.33
Item ($\hat{\tau}_{00(item)}$)						
Intercept	30.04			0.0716		
Residual ($\widehat{\sigma^2}$)	163.07			0.3301		

Table 5. Parameter estimates from LMMs fit to raw and inverse-transformed SPP data. These estimates were used as the input parameters for the simulations performed in Studies 1.2 and 1.3 respectively.

Study 1.2: Simulations based on raw model as generating process

Generating process. The parameters used in the generating process of this simulation were integer-rounded values of the significant fixed-effect estimates (implying 0 for x_2z_1 and $x_1x_2z_1$) and of the random-effect estimates obtained from the raw LMM (summarized in Table 5). Because the raw LMM assumed normally distributed residuals, its residual variance estimate σ^2 cannot be directly used to generate trial-level residuals. Instead, to mimic the skew and variance observed in the raw RT data, a Gamma distribution with shape parameter = 0.15 and scale parameter = 1000 was used to generate trial-level residuals. Finally, an intercept parameter of 600 was added to the simulated values to approximate the intercept estimated by the raw model.

Dataset generation and analysis. The datasets were generated to simulate a typical semantic priming experiment. In this experiment, all subjects see the same targets, but half the targets are preceded by related primes. Prime relatedness (x_2) was counterbalanced in two lists so that targets preceded by related primes in one list are preceded by unrelated primes in the other list. Based on their distributions in

the megastudy, word frequency (x_1) and vocabulary (z_1) values were generated from standard normal distributions and these values were sorted so that word frequencies are matched between the relatedness conditions.

The datasets were set up to have 20 to 200 subjects fully crossed with 20 to 200 items in increments of 20, resulting in 100 sample size conditions that more than encompass the typical sizes of psycholinguistic studies; one thousand datasets were generated for each of the 100 conditions. In each dataset, a random subset of up to 5% of all the observations were deliberately excluded to simulate incorrect responses and missing data. All datasets were then finally analyzed using the same preprocessing and fitting procedure described in the Study 1.1.

A new dataset was generated whenever a warning or an error about convergence arose in fitting the raw model. This procedure was repeated for each sample size condition until there were 1,000 datasets where the raw model seamlessly converged.

Performance measures: Raw model. The consequences of violating the normality assumption by fitting LMMs to raw RT data were assessed by calculating percentage bias and coverage for each effect in the model. The effects of subject and item sample sizes on these performance measures were also assessed. Percentage bias was computed as [(estimate $\hat{\gamma}$ – true value γ)/true value γ] $\times$ 100, where positive values indicate that the parameter was overestimated and values between 5-10% are considered tolerable (e.g., Kaplan, 1989). For null/non-existing effects, bias was simply the estimate $\hat{\gamma}$. Coverage was determined as the percentage of occasions that the 95% confidence interval for estimate $\hat{\gamma}$ included the true value γ.[5] Coverage provides a measure of whether the confidence intervals generated from the model tend to be conservative or permissive/anti-conservative, where percentages below 95% indicate conservative coverage.

Comparison measures: Transformed vs. raw model. Because the four RT scales consist of different units of measurement, our ability to compare the models fit to each RT scale was limited.

[5]The critical *t* value of |2| was also used to compute the 95% confidence intervals for all the parameter estimates.

Moreover, models that converge when fit to raw RT do not necessarily converge when fit to the transformed data and vice versa. Therefore, the models were first compared in terms of how consistent they were in converging. Subsequent comparisons were limited to datasets where both the raw model and its transformed counterpart both converged.

The two models were then compared on how much power they had in detecting effects present in the simulation and how much Type I error they incur for null effects/effects absent in the simulation (i.e., x_2z_1 and $x_1x_2z_1$), where power (Type I error) is defined as the proportion of models that meet the *t* critical value of $|2|$, when said effect is present (absent) in the simulation. To model how changes in subject and item sample sizes affect these measures, I fit loess curves to the power and Type I error estimates from all the sample size conditions and plotted these curves onto contour plots. To model how *differences* in power/Type I error between the transformed and raw models change due to sample size, I subtracted the estimates obtained from the raw and transformed LMMs at each sample size condition, fit loess curves to these difference estimates, and plotted these difference curves onto contour plots.

Results: Study 1.2

Performance measures: Raw model.

***Bias and coverage*.** Figure 3 shows the average percentage bias for each effect in the raw model as a function of subject and item sample size. Averaging across all subject and item sample sizes, the raw model underestimated all the existing effects in the model: the x_1, x_2, and z_1 main effects by 5.99%, 0.34%, and 7.91% respectively; and the x_1x_2 and x_1z_1 interaction effects by 5.84% and 7.42% respectively. The average bias for the null effects x_2z_1 and $x_1x_2z_1$ were negligible. Lastly, other than volatile estimates in models fit to small sample sizes, changes in sample size was not related to how much bias is incurred by the raw model.

Figure 4 shows the coverage for each effect in the raw model as a function of subject and item sample size. The raw model produced slightly conservative coverage for x_1: averaging across all subject and item sample sizes, only 93.1% of the generated 95%-confidence intervals contained the true value of

x_1. Moreover, increasing both subject and item sample sizes appeared to lower the coverage for this effect. All other effects had coverages that approximated the nominal 95% value. For these effects, changes in sample size did not affect their coverage estimates.

***Power/Type I error*.** The contour plots in Figure 5 show predicted changes in power/Type I error for the raw model as a function of subject and item sample sizes. As expected, power for all non-zero effects increased as subject and item sample sizes increased. Notably, some effects are estimated to be more powerful than others: for instance, x_1 already has massive power at 0.80 with 50 subjects and 50 items, whereas x_2's estimated power is only at 0.50 and z_1, x_1x_2, and x_1z_1's power estimates are below 0.20 with these sample sizes. Lastly, violating the normality assumption did not seem to increase Type I error rates for the null effects: regardless of subject and item sample size, Type I error estimates for x_2z_1 and $x_1x_2z_1$ remained below 7%.

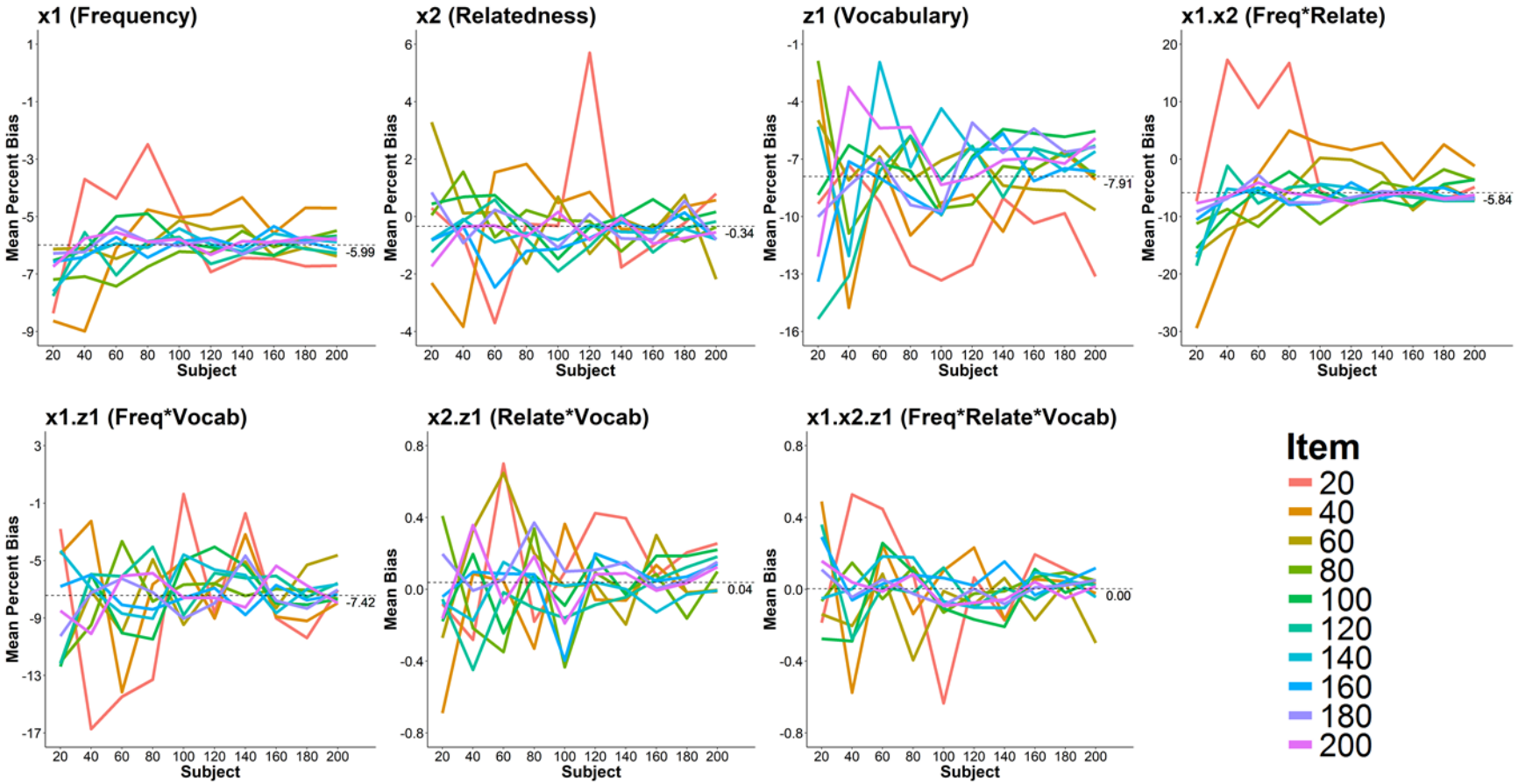

Figure 3. Average bias for each of the fixed-effect estimates in the raw model as a function of subject and item sample sizes. Simulation based on the model fit to the raw SPP data.

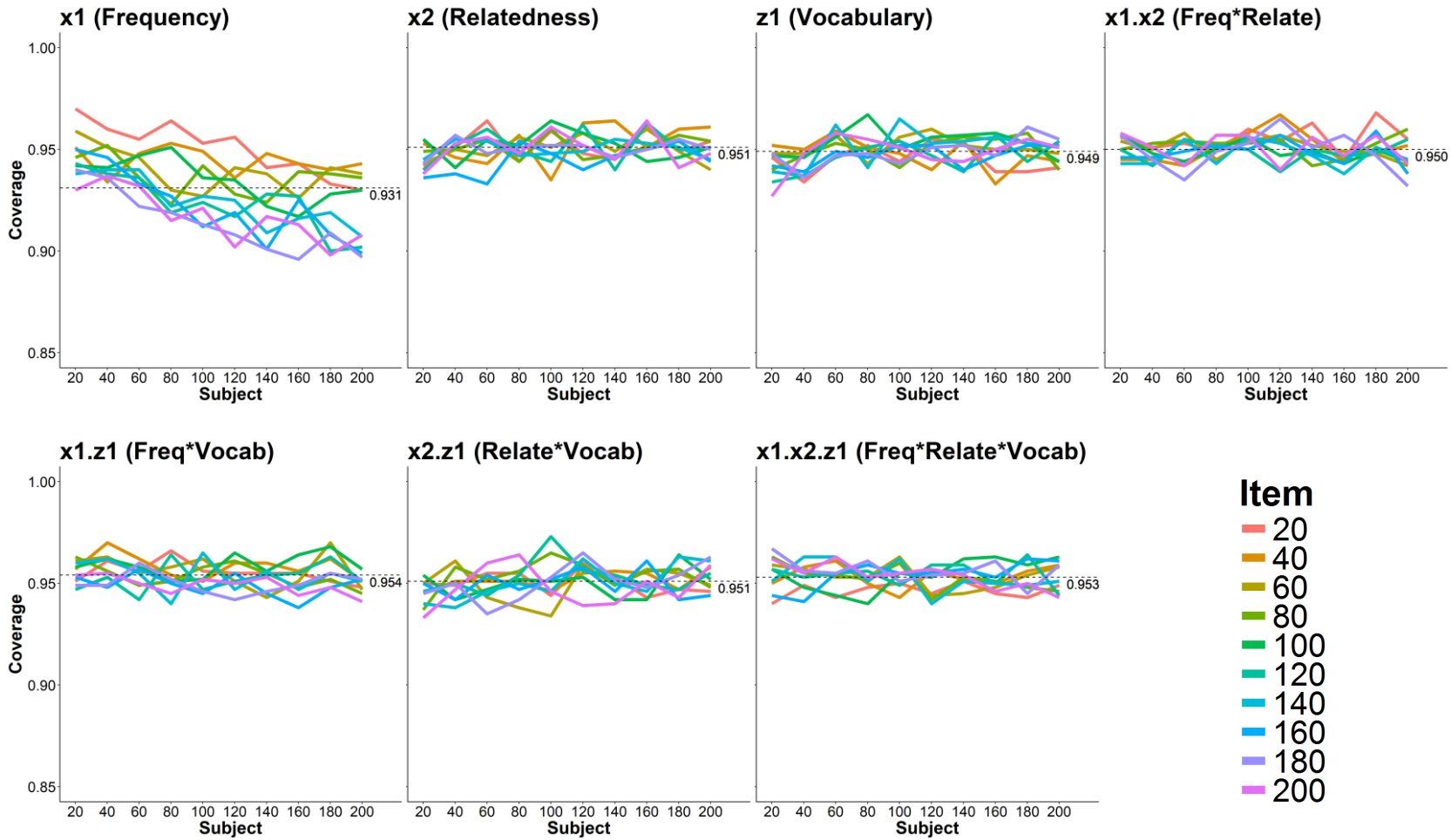

Figure 4. Coverage for each of the fixed effects in the raw model as a function of subject and item sample sizes. Simulation based on the model fit to the raw SPP data.

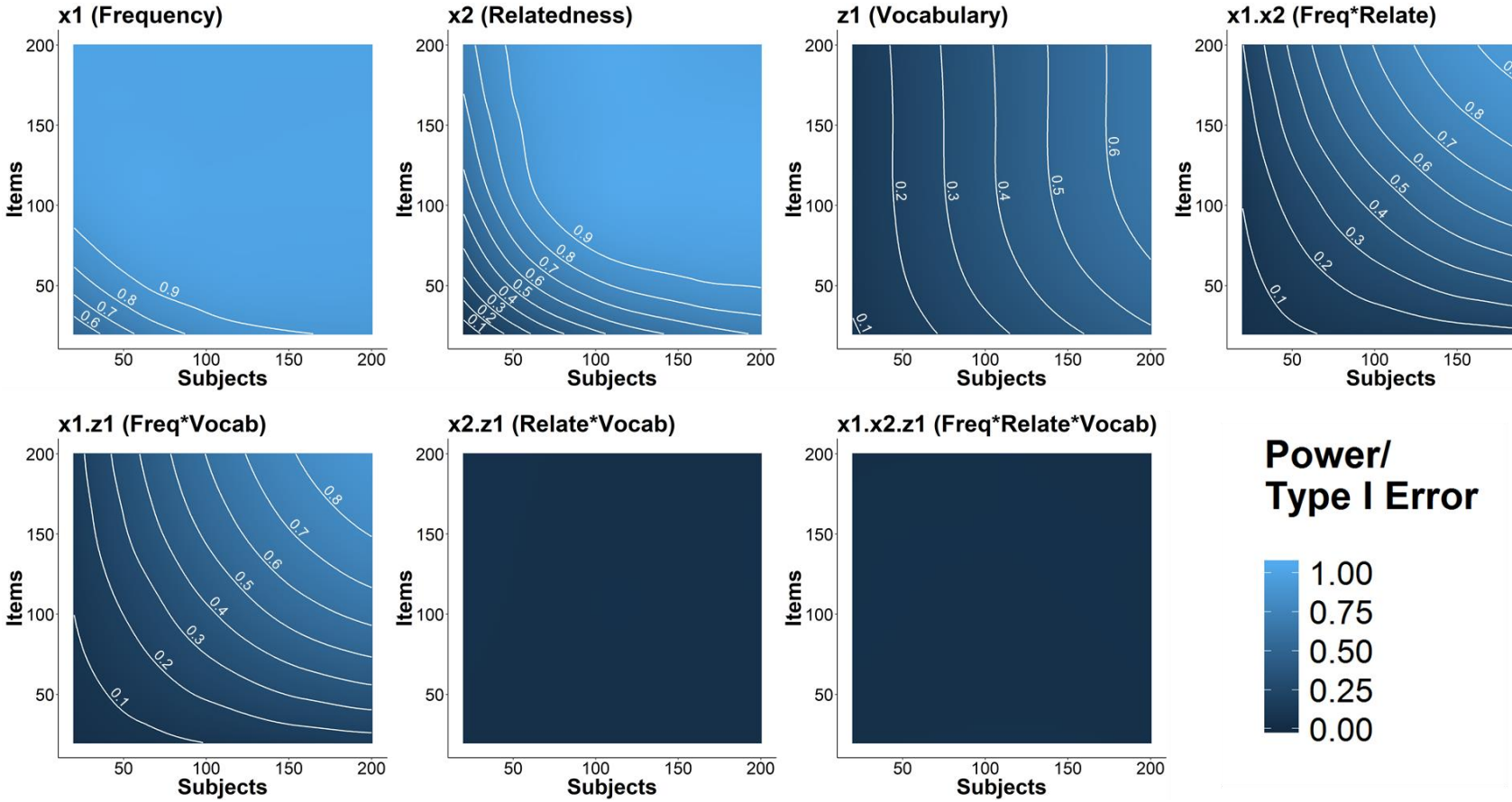

Figure 5. Power (Type I error) contour plots for the raw model as a function of subject and item sample size. The contour plots for $x_2 z_1$ and $x_1 x_2 z_1$ are Type I error contour plots; the rest are power contour plots. Simulation based on the model fit to the raw SPP data.

Comparison measures: Transformed vs. raw model. Figure 6 shows the proportion of datasets where both the raw and transformed model converged as a function of subject and item sample sizes. Clearly, a model fit to the raw data did not necessarily converge when fit to the transformed data. Moreover, item sample sizes appear to have a greater effect on this convergence consistency, such that when the sample had fewer items, models that converged when fit to the raw data tended not to converge when fit to the transformed data.

Subsequent comparisons between the raw and the transformed models were then made on datasets where both the raw and transformed model converged.

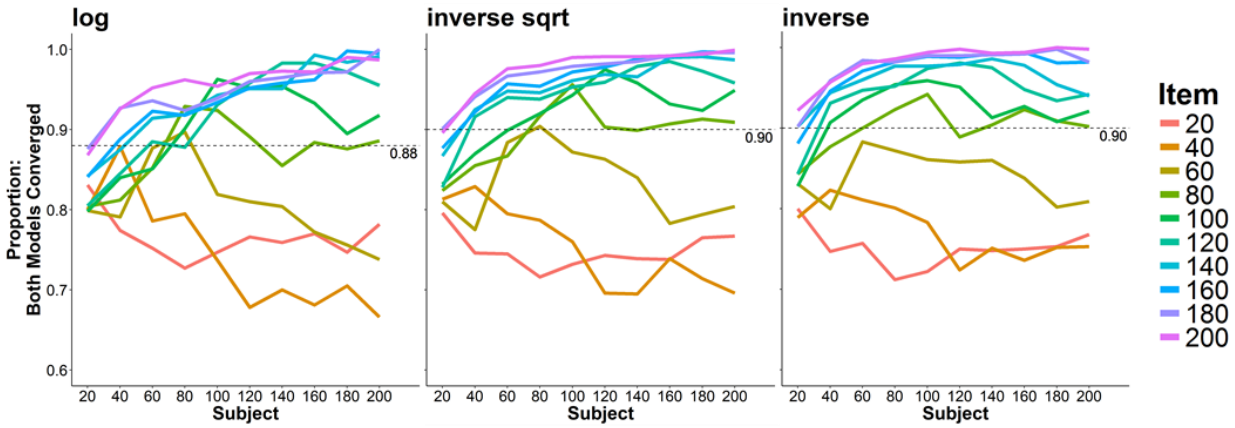

Figure 6. Proportion of datasets where the transformed models converged given that the raw model also converged, expressed as a function of subject and item sample sizes. Simulation based on the model fit to the raw SPP data.

Log model. The log model converged in 87,814 of the 100,000 datasets where the raw model converged. As in the raw model, increasing subject and item sample sizes also increased power in the log model for all the existing effects as shown in Figure 7. Moreover, Type I error rates for the null effects still remained below 7% regardless of sample size.

Next, I calculated the difference in power/Type I error between the raw and the log model for each of the effects. I then fit loess curves onto the differences using the subject and item sample sizes as predictors. The contour plots in Figure 8 show the results of these models. For x_1 and x_2, the log model was estimated to have as much as 17% and 21% more power than the raw model given smaller sample

sizes respectively. For the x_1x_2 and x_1z_1 interactions, the log model was estimated to have as much as 7% to 15% more power as sample sizes increased respectively. Differences in power for z_1 and Type 1 error for x_2z_1 and $x_1x_2z_1$ remained below 5% across all sample sizes.

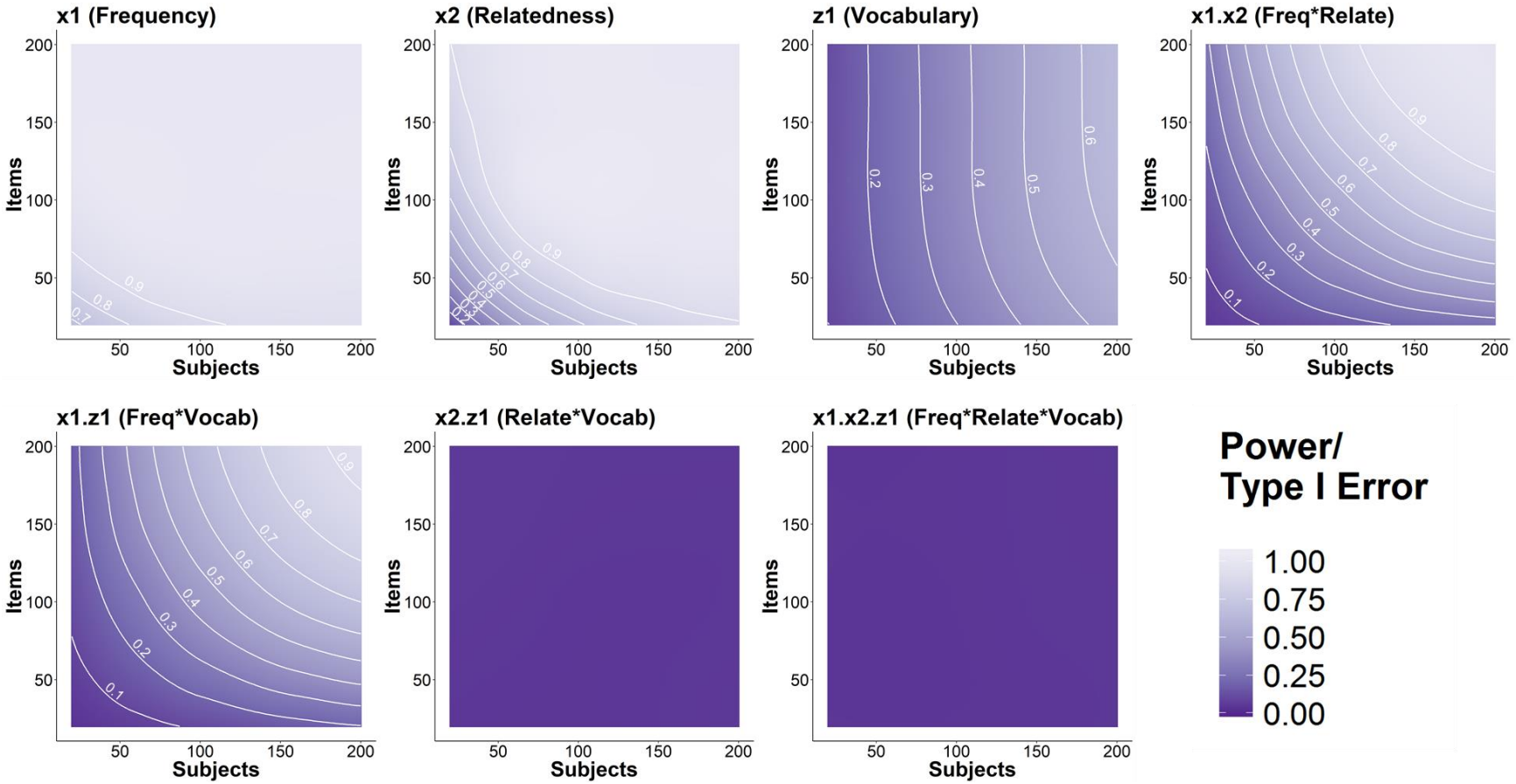

Figure 7. Power (Type I error) contour plots for the log model as a function of subject and item sample size. The contour plots for x_2z_1 and $x_1x_2z_1$ are Type I error contour plots; the rest are power contour plots. Simulation based on the model fit to the raw SPP data.

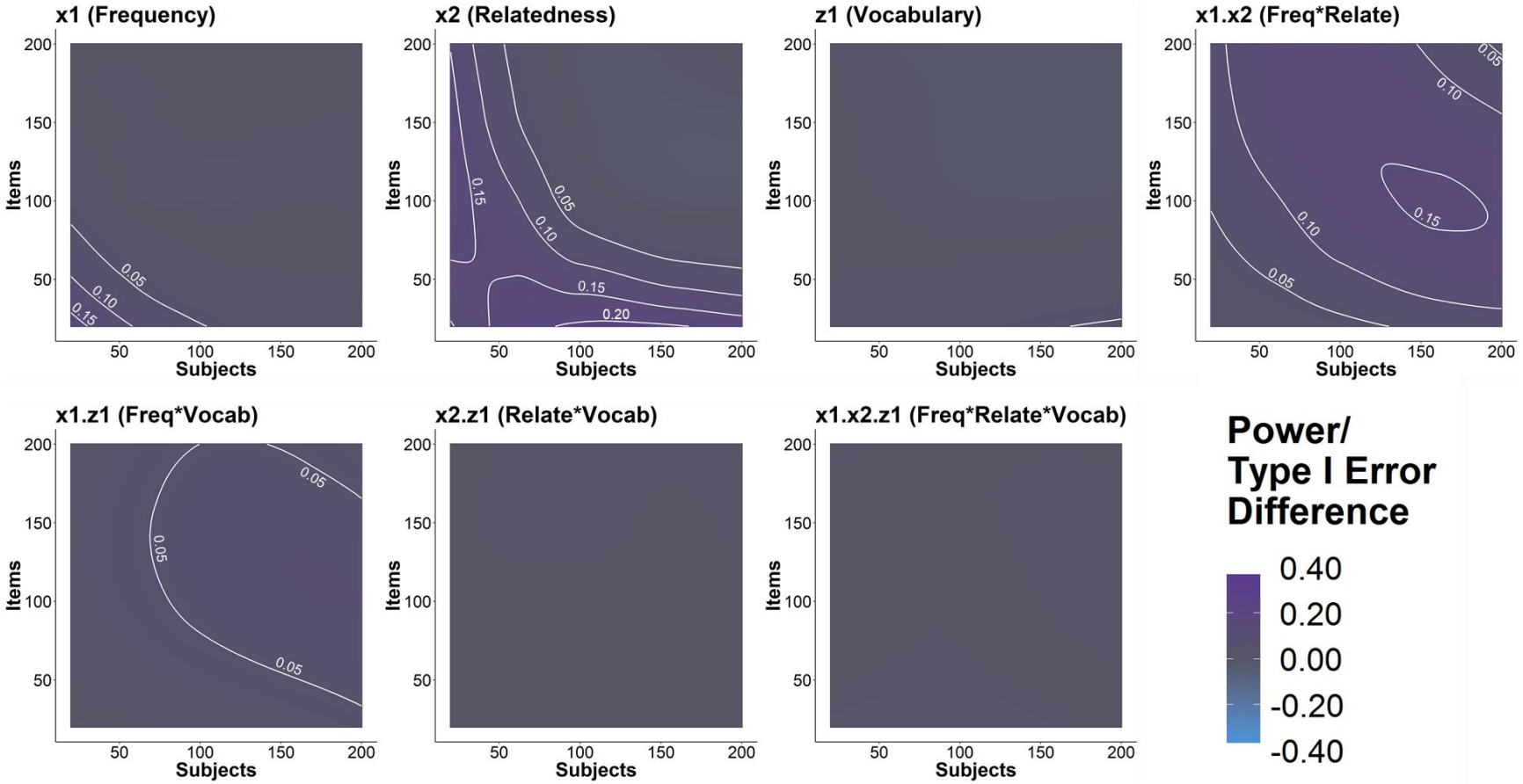

Figure 8. Power (Type I error) difference contour plots (log – raw) as a function of subject and item sample size. The contour plots for $x_2 z_1$ and $x_1 x_2 z_1$ are a Type I error contour plots; the rest are power contour plots. Blue indicates that raw model has more power (Type I error) than the log model. Purple indicates that log model has more power (Type I error) than the raw model. Simulation based on the model fit to the raw SPP data.

Inverse square-root model. The inverse square-root model converged in 89,568 of the 100,000 datasets where the raw model converged. Figure 9 shows the general increase in power for all existing effects as sample size increased, whereas Type I error rates for the null effects remained below 8% regardless of sample size.

Figure 10 shows the estimated differences in power/Type I error between the raw and the inverse square-root model for each of the effects as a function of subject and item sample sizes. For x_1, x_2, and z_1, the inverse square-root model was estimated to have as much as 22%, 29%, and 8% more power than the raw model given smaller sample sizes respectively. This indicates that the inverse square-root model has more power for the main effects than the log and raw models in smaller sample sizes. As for interactions, the inverse square-root model had greater power over the raw model only in the x_1x_2 effect, where it had as much as 17% more power than the raw model as sample sizes increased. Differences in power for x_1z_1 and in Type I error for x_2z_1 and $x_1x_2z_1$ between the raw and inverse square-root model remained below 5% across sample sizes.

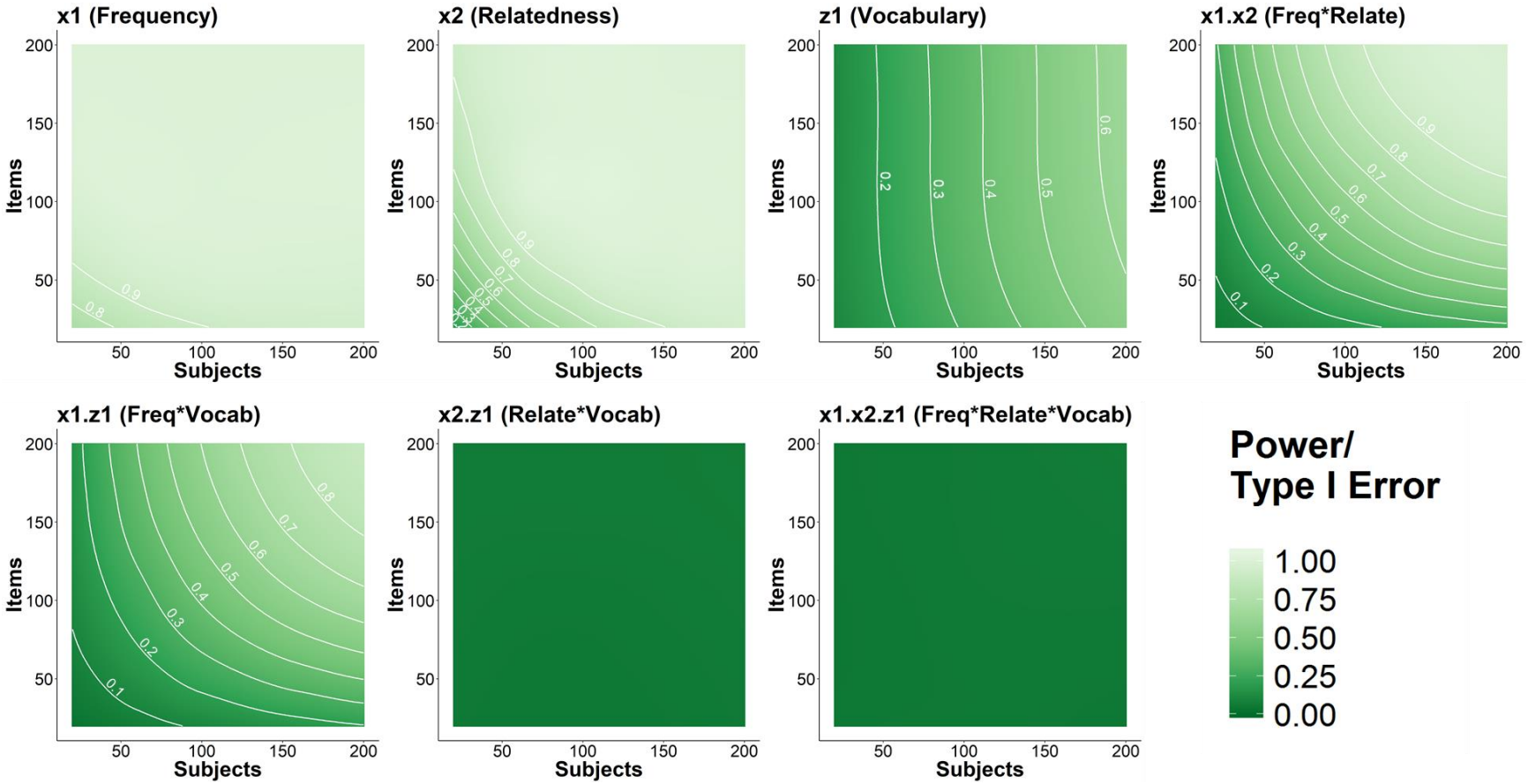

Figure 9. Power (Type I error) contour plots for the inverse square-root model as a function of subject and item sample size. The contour plots for x_2z_1 and $x_1x_2z_1$ are Type I error contour plots; the rest are power contour plots. Simulation based on the model fit to the raw SPP data.

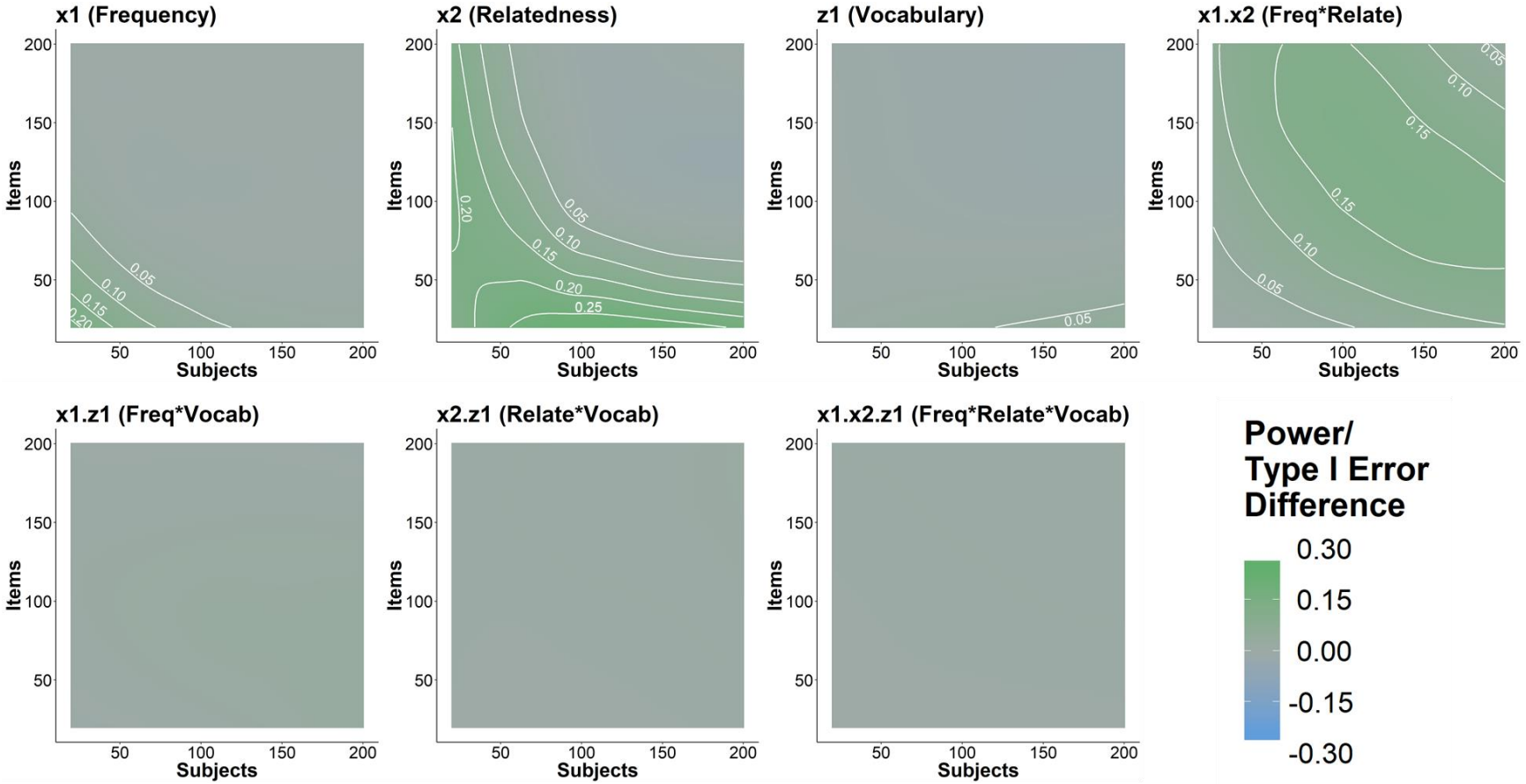

Figure 10. Power (Type I error) difference contour plots (inverse square-root – raw) as a function of subject and item sample size. The contour plots for x_2z_1 and $x_1x_2z_1$ are a Type I error contour plots; the rest are power contour plots. Blue indicates that raw model has more power (Type I error) than the inverse-square root model. Green indicates that inverse-square root model has more power (Type I error) than the raw model. Simulation based on the model fit to the raw SPP data.

Inverse model. The inverse model converged in 90,181 of the 100,000 datasets where the raw model converged. As in all the other model types, Figure 11 shows the general increase in power for all existing effects as sample size increased, whereas Type I error rates for the null effects remained below 10% regardless of sample size.

Figure 12 shows the estimated differences in power/Type I error between the raw and the inverse model for each of the effects as a function of subject and item sample sizes. For x_1, x_2, and z_1, the inverse model was estimated to have as much as 24%, 36%, and 9% more power than the raw model given smaller sample sizes respectively. This indicates that the inverse model is the most powerful model of the four model types for the main effects given smaller sample sizes. As for interactions, the inverse model had greater power over the raw model only in the x_1x_2 effect, where it had as much as 15% more power than the raw model as sample sizes increased. However, the raw model notably had greater power than the inverse model for the x_1z_1 effect, where it had as much as 12% more power than the inverse model as sample sizes increased. Differences in Type I error rates for x_2z_1 and $x_1x_2z_1$ between the raw and inverse model remained below 5% regardless of sample size. Recall that the inverse transform optimally normalized the real RT data whose analysis was the basis for this simulation.

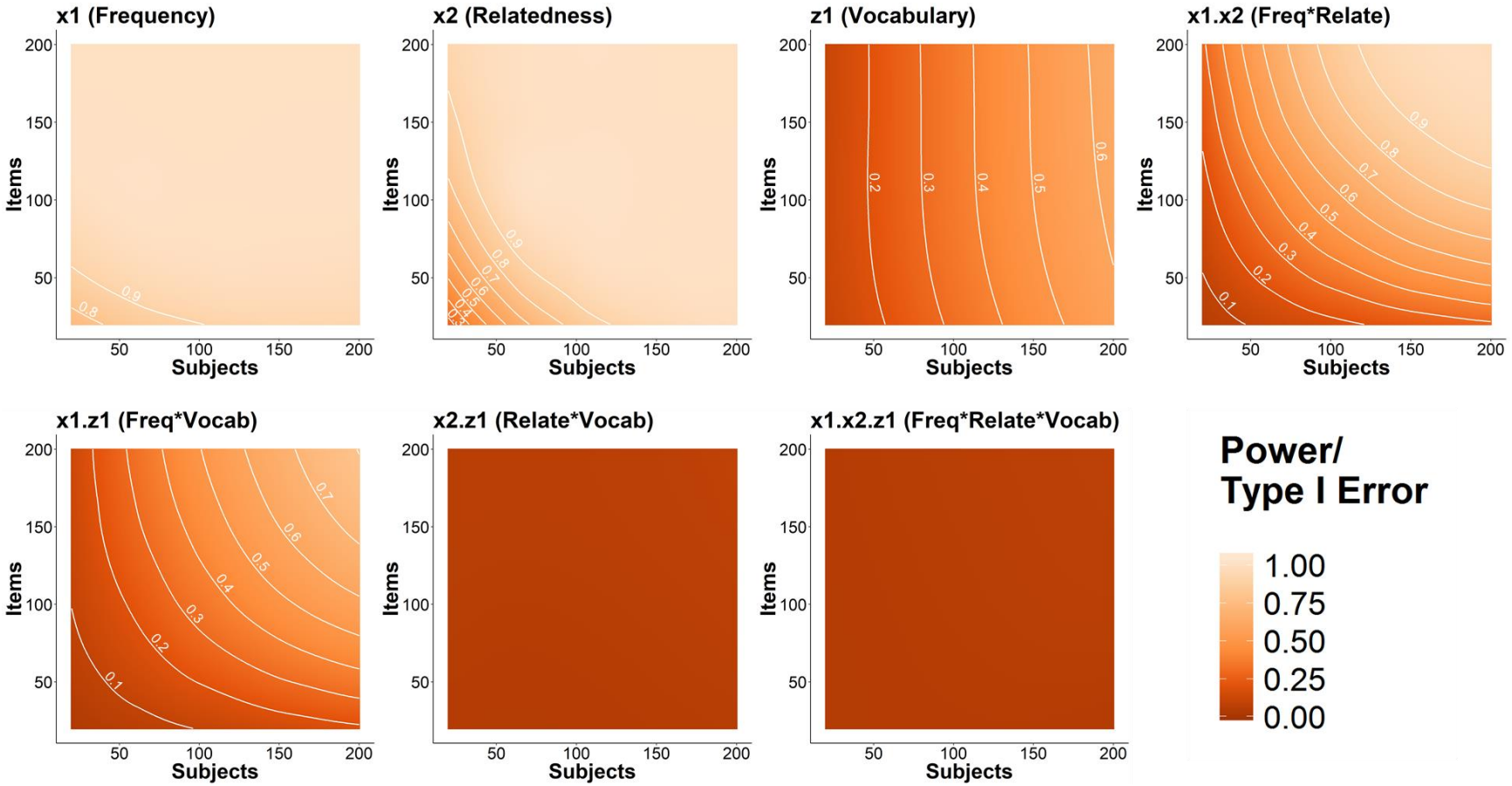

Figure 11. Power (Type I error) contour plots for the inverse model as a function of subject and item sample size. The contour plots for $x_2 z_1$ and $x_1 x_2 z_1$ are Type I error contour plots; the rest are power contour plots. Simulation based on the model fit to the raw SPP data.

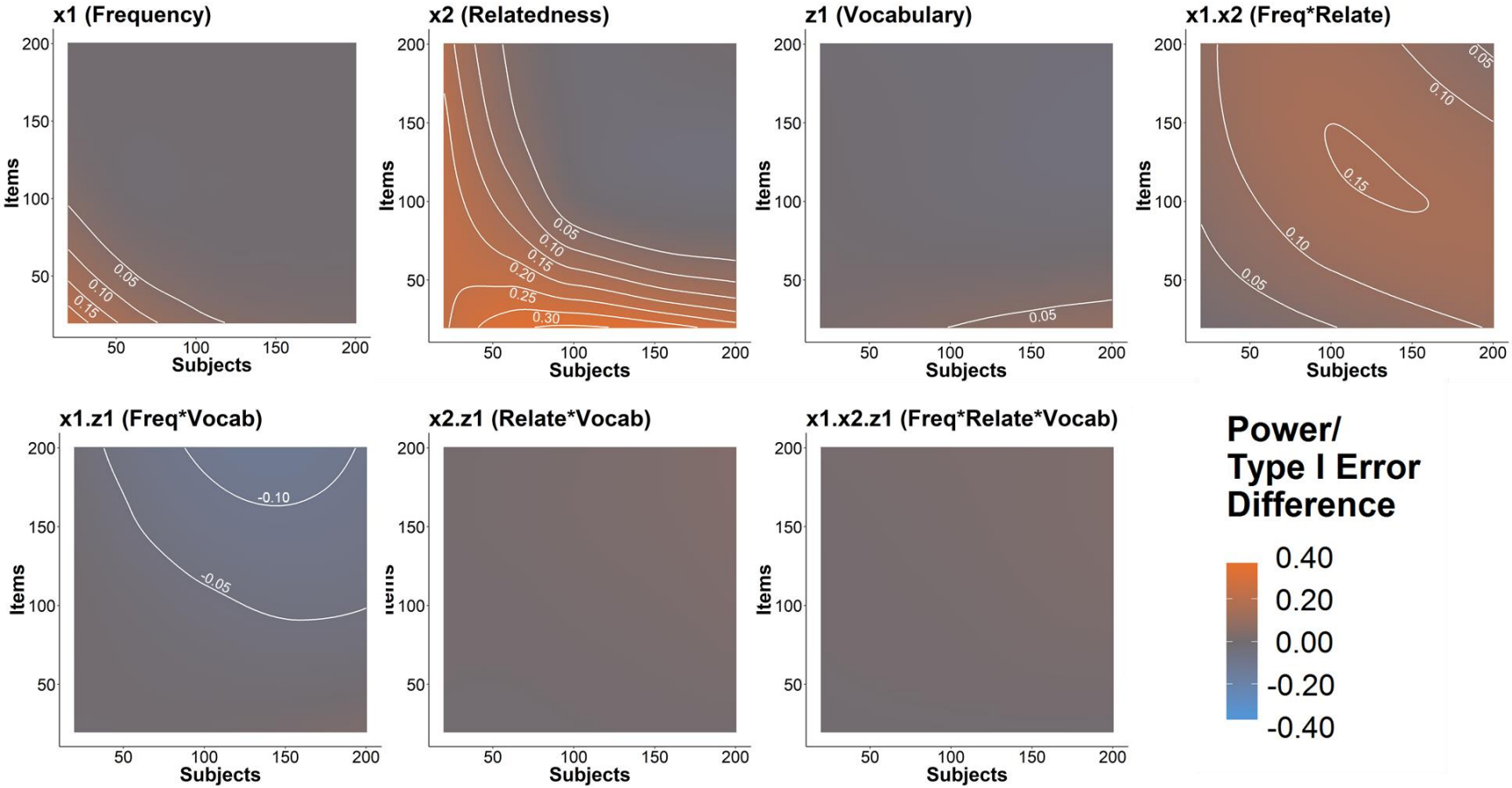

Figure 12. Power (Type I error) difference contour plots (inverse – raw) as a function of subject and item sample size. The contour plots for x_2z_1 and $x_1x_2z_1$ are Type I error contour plots; the rest are power contour plots. Blue indicates that raw model has more power (Type I error) than the inverse model. Orange indicates that inverse model has more power (Type I error) than the raw model. Simulation based on the model fit to the raw SPP data.

Summary. There were three main findings from this simulation. First, violating the normality assumption by fitting a standard LMM to raw data resulted in underestimated fixed effects and more conservative confidence intervals for one main effect (x_1). Second, across all RT scales, increasing sample sizes increased power for existing effects, but it did not affect Type I error rates for null effects.

Lastly, as the power transform became stronger, the model became more powerful in detecting main effects at smaller sample sizes, but the results for detecting interactions as sample sizes increased were mixed. Transformed models were more powerful than the raw model in detecting x_1x_2 as sample sizes increased. However, the opposite held for x_1z_1: as the power transform became stronger, the power advantage for transformed models disappeared, and the raw model became more powerful in detecting x_1z_1 as sample sizes increased. Negligible differences in Type I error for the null effects were observed across the model types, and these were not affected by changes in sample size.

Despite underestimating the main effects, the raw model was less powerful than the transformed models in detecting them only at smaller sample sizes. More surprising is that despite underestimating the x_1z_1 interaction effect, the raw model was still *more* powerful than the inverse model in detecting it as sample sizes increased. This is despite the SPP analysis showing that the inverse model optimally normalized the residuals from the raw model (i.e., the model used to generate this simulation).

It is possible that these results were obtained because the simulation was set up with the raw LMM as the basis of the data-generating process. This concern was addressed in Study 1.3 by simulating data using the inverse LMM as the basis of the data-generating process. If the generating process of the simulations were to be based on the (optimally) transformed model, the transformed model's performance over the raw model might be even better than when the generating process was based on the raw model.

Study 1.3: Simulations based on inverse model as generating process

Generating process. The parameters used in the generating process of this simulation are the significant fixed-effect estimates (implying 0 *only for* $x_1x_2z_1$) and random-effect estimates obtained from the inverse model (summarized in Table 5). With the Box-Cox procedure revealing that the inverse

transform optimally normalizes the RT data, a normal distribution with mean = 0 and variance = 0.3301^2 = 0.1090 was used to generate trial-level residuals.

Dataset generation and analysis. This was identical to Study 1.2, except that the simulated inverse RTs were back-transformed only into raw RTs to compare the performance of the models fit to these two RT scales.

Comparison measures: Inverse vs. raw model. The same comparison measures were obtained as in Study 1.2.

Results: Study 1.3

Figure 13 shows the proportion of datasets for which both the raw and inverse models converged as a function of subject and item sample sizes. As in Study 1.2, models that converged when fit to the raw data did not necessarily converge when fit to the inverse-transformed data, and smaller item sample sizes seemed to exacerbate this issue. Subsequent results are evaluated on the 88,379 datasets where inverse model also converged, out of 100,000 datasets where the raw model converged.

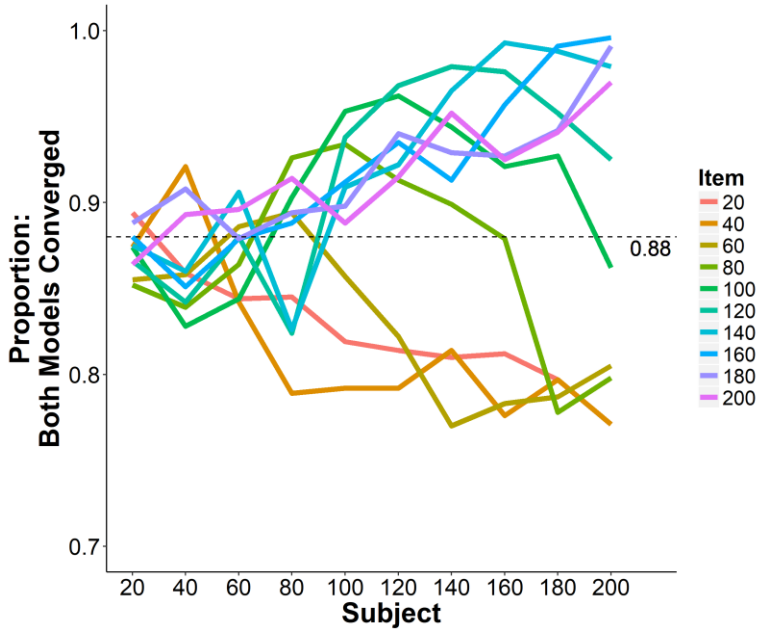

Figure 13. Proportion of datasets where the inverse models converged given that the raw model also converged, expressed as a function of subject and item sample sizes. Simulations based on the model fit to inverse-transformed SPP data.

The power contour plots for the raw model and inverse model in Figures 14 and 15 show the general increase in power as subject and item sample sizes increase. The only difference between these contour plots is that increasing sample sizes increased power for x_2z_1 in the inverse model but not in the raw model. Type I error for the three-way interaction remained below 10% for both the raw and inverse model.

Differences between the raw and inverse models' power estimates as a function of sample sizes are seen in the power difference contour plots in Figure 16. The inverse model's power for the x_1 and x_2 main effects was as much as 8% and 14% higher than the raw model's in smaller sample sizes respectively, but this difference disappeared as sample sizes increased. In contrast, the results for the interactions are mixed. For the x_1x_2 and x_1z_1 interactions, the raw model had as much as 13% and 25% more power than the inverse model as sample sizes increased respectively. In contrast, for the x_2z_1 interaction, the inverse model had as much as 10% more power than the raw model as sample sizes increased. Differences in power for the z_1 effect and in Type I error for the three-way interaction between the raw and inverse model remained lower than 5% across sample size conditions.

Some of the results were similar to those observed in Study 1.2: the inverse model had more power for the main effects in smaller sample sizes, the raw model had more power in detecting the x_1z_1 effect as sample sizes increased, and Type I error for the three-way-interaction remained at/below 5% regardless of sample size for both models. But some results were different: while the inverse model showed more power for the x_1x_2 effect when the generating process was based on the raw model, the raw model showed *more* power for the same effect when the generating process was based on the *inverse model*. Moreover, the inverse model showed greater power for the x_2z_1 effect in this study, but it did not exhibit greater Type I error for this effect in Study 1.2. Overall, these results are inconclusive about whether the choice of underlying generating process affects the influence of power transforms on the fixed effects.

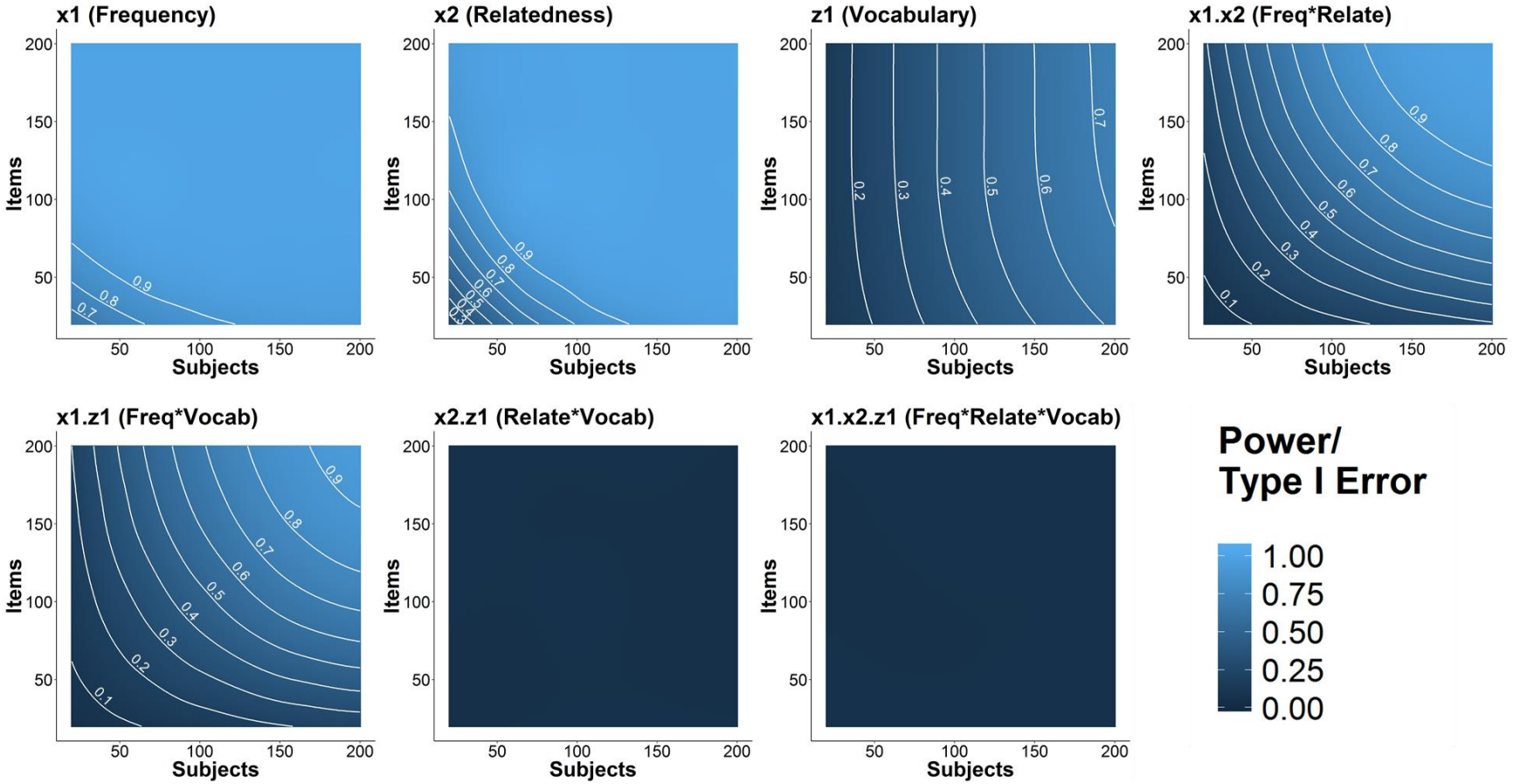

Figure 14. Power (Type I error) contour plots for raw model as a function of subject and item sample size. The contour plot for $x_1x_2z_1$ is a Type I error contour plot; the rest are power contour plots. Simulation based on the model fit to the inverse-transformed SPP data.

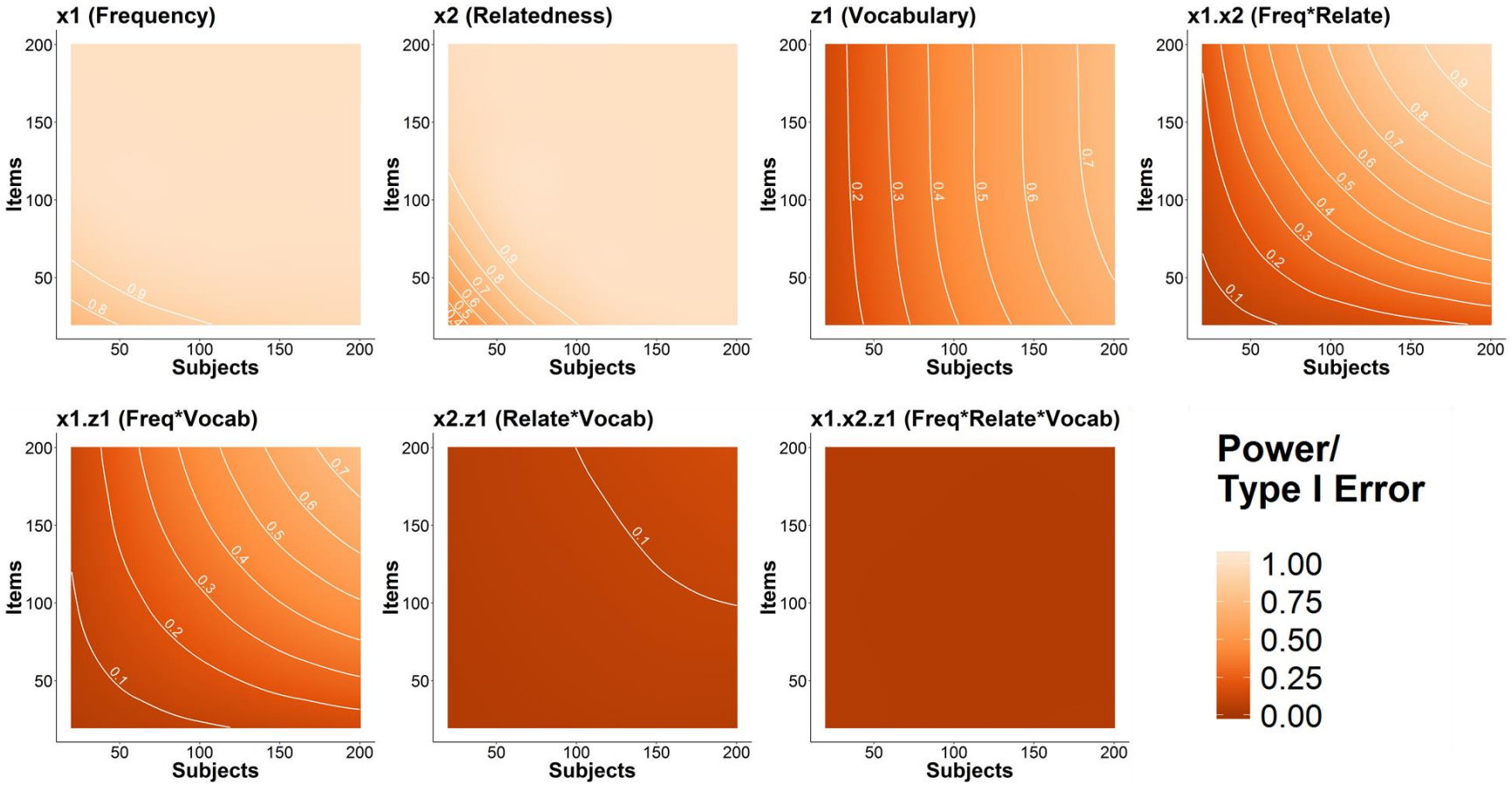

Figure 15. Power (Type I error) contour plots for inverse model as a function of subject and item sample size. The contour plots for $x_1x_2z_1$ is a Type I error contour plot; the rest are power contour plots. Simulation based on the model fit to the inverse-transformed SPP data.

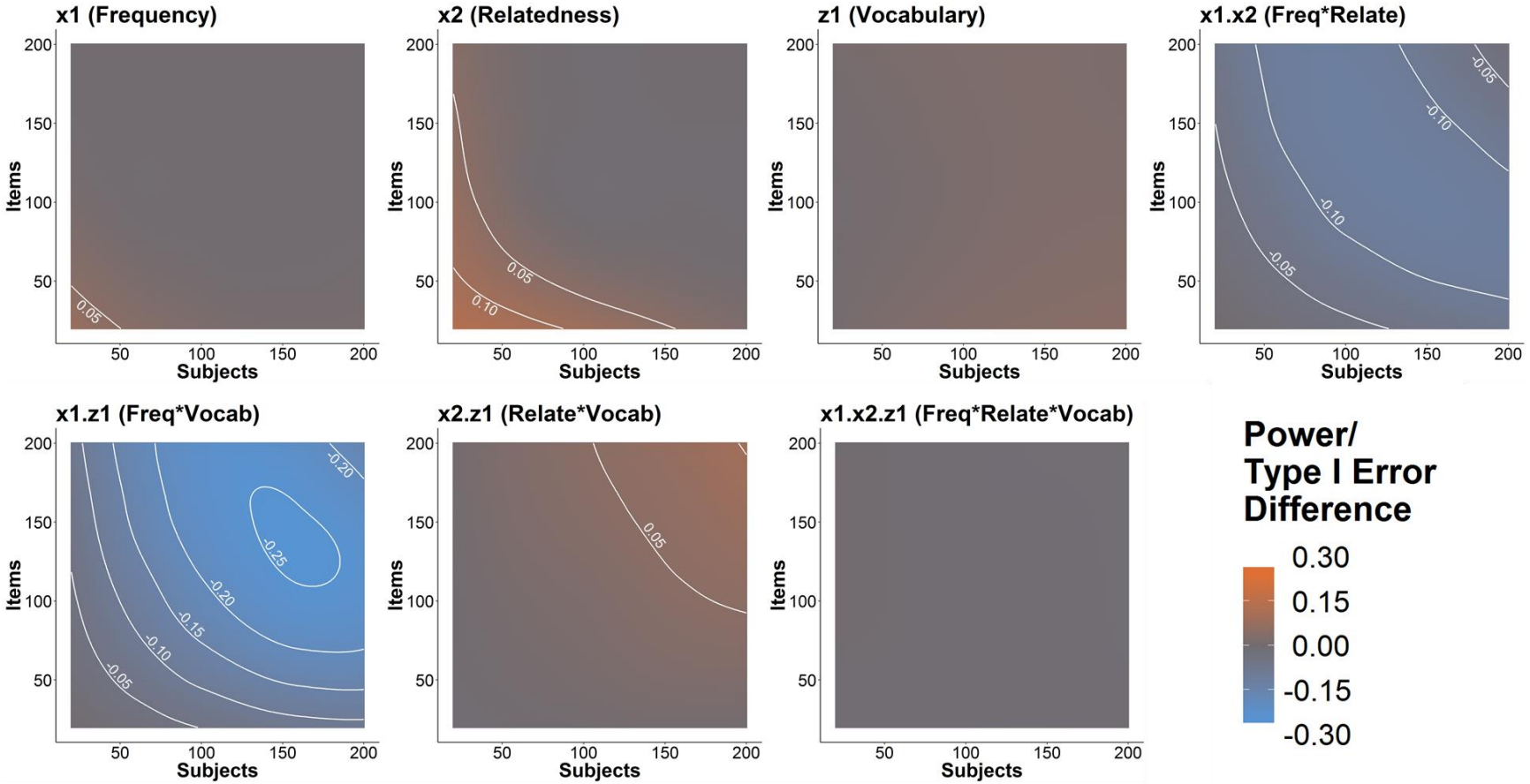

Figure 16. Power (Type I error) difference contour plots (inverse – raw) as a function of subject and item sample size. The contour plots for $x_1x_2z_1$ is a Type I error contour plot; the rest are power contour plots. Blue indicates that raw model has more power (Type I error) than the inverse model. Orange indicates that inverse model has more power (Type I error) than the raw model. Simulation based on the model fit to the inverse-transformed SPP data.

Discussion

Analysis of the SPP revealed that power transforms led to systematic changes in the *t*-statistics of the fixed effects. Stronger power transforms slightly increased the *t*-statistics of the main effects, decreased those of x_1x_2 and x_1z_1 effects, and increased that of x_2z_1 to significance. They also altered random effect correlation patterns present in the raw scale.

These results were partially supported by the simulations. Analogous to the slight increases in the t-statistics of the main effects as stronger power transforms were applied to the SPP data, the simulations revealed that the transformed LMMs generally had greater power than the raw LMM for main effects in small sample sizes. This is consistent with results from simulations performed on smaller sample sizes in the context of ANOVAs (Levine & Dunlap, 1982; Ratcliff, 1993). Moreover, analogous to the systematic decrease in the *t*-statistic of the x_1z_1 interaction as stronger power transforms were applied to the SPP data, the LMMs also became less sensitive in detecting the x_1z_1 interaction as stronger power transforms were applied in the simulations. However, some results seemed to depend on the generating process of the simulation. In Study 1.2 where the raw LMM was the basis of the generating process, the transformed model was more sensitive in detecting the x_1x_2 interaction than the raw model as sample sizes increased. But in Study 1.3 where the inverse LMM was the basis of the generating process, the raw LMM surprisingly outperformed the inverse LMM in detecting the x_1x_2 interaction as sample sizes increased.

The ambiguous results from the simulations based on the SPP motivate the analysis of other megastudies and the replication of the simulations using different data-generating processes to determine whether the results obtained here generalize to other word recognition experiments. This goal is pursued in the following two studies.

Study 2: Form Priming Project (FPP)

The FPP (Adelman et al., 2014) is a megastudy of the orthographic priming phenomenon in word recognition. In a typical masked-form orthographic priming paradigm, a mask (e.g., ######) is presented for ~500 ms, followed briefly (~50 ms) by an orthographic prime before the target word (e.g., doctor) is

presented. Responses tend to be faster on average for targets preceded by orthographically related primes (i.e., primes that targets share letters with; e.g., odctro) than unrelated/baseline primes (e.g., xpqalb) (Forster, Davis, Schoknecht, & Carter, 1987). The FPP consists of data from 1,015 subjects, each of whom responded to 420 prime-target word pairs and took a vocabulary and spelling test. More details about the procedure are available in Adelman et al. (2014).

In Study 2, I focused on the following three predictors in the FPP and their interactions:

1) The target word frequency (x_1), which is identical to the frequency measure in Study 1;
2) Type of orthographic prime (x_2), which is a word-level manipulation indicating whether the target was preceded by an orthographically related or baseline prime, thus representing the orthographic priming effect[6]; and
3) Vocabulary (z_1), which is a continuous person-level characteristic representing the score of a subject on a multiple-choice vocabulary test

Study 2.1: Analyzing FPP Data

Data preprocessing. Only RTs from accurate responses to lexical targets (i.e., word strings correctly identified as words) were analyzed. Following Adelman et al. (2014), data from 43 subjects were excluded; these subjects came from testing sites that had equipment timing issues and whose accuracy was below 75%. Other than these steps, the same preprocessing procedure in Study 1 was applied to the FPP data, only replacing x_2 as orthographic prime type. This preprocessing procedure resulted in 1.46% to 3.12% of observations being dropped from further analyses across RT scales. Four subjects were further excluded due to missing vocabulary scores, resulting in 958 to 960 subjects being retained for analyses.

[6]There were 28 priming conditions in the FPP. For the LMM analysis, I categorized 26 of these conditions as "orthographically related" conditions and the other two as baseline conditions (i.e., where the prime does not share any letters with the target). This dichotomy considerably simplified the analysis strategy but left unequal sample sizes for the orthographically related condition (n = 390) and the baseline condition (n = 30). Unequal sample sizes may lead to heteroscedastic errors and loss of power, particularly for small sample sizes. However, because the FPP dataset is massive, this inequality is less consequential.

Fitting the LMMs and identifying the optimal power transform. The main effects of word frequency, type of prime, and vocabulary and all interaction effects were entered into a cross-classified LMM, one for each RT scale. I specified random intercepts for subjects and items and random subject slopes for the word frequency (x_1) and prime type (x_2) effects.

After the models were fit, the optimal power transform for the preprocessed data was identified as in Study 1. The Box-Cox procedure revealed that the preprocessed data is optimally normalized using λ = -0.71, which in strength is closer to the inverse square-root transform than the inverse transform. The QQ plots in Figure 17 show that while the inverse transform most reduced the residuals' positive skew, it also introduced the most negative skew among the transforms. The Box-Cox procedure therefore suggests that the optimal transform balances the reduction of positive skew with the potential introduction of negative skew. In this case, the optimal transform is more similar to the inverse square-root transform.

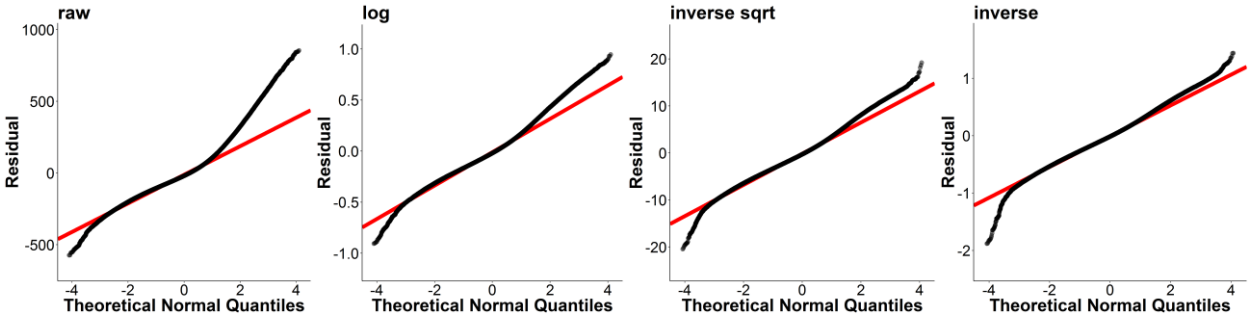

Figure 17. Trial-level residual QQ plots of models fit to FPP data

The *t*-statistics across all four models showed a similar pattern to that observed in the SPP analysis, although to a less extent. Table 6 shows that while none of the effects' significance changed, the *t*-statistics for x_2 increased and *t*-statistics for all the two-way interactions shrank as a stronger power transform was applied to the data.

RT Scale	x_1 (Freq)	x_2 (Prime)	z_1 (Voc)	x_1x_2 (Freq x Prime)	x_1z_1 (Freq x Voc)	x_2z_1 (Prime x Voc)	$x_1x_2z_1$ (F x P x V)
Raw	-19.69	-20.67	-3.70	4.67	4.04	1.42	0.32
Log	-19.68	-24.89	-3.68	3.42	3.31	0.56	0.86
Inverse Sqrt	-19.46	-26.14	-3.74	2.59	3.26	0.44	0.57
Inverse	-19.32	-27.36	-3.82	2.18	2.84	-0.09	0.67

Table 6. *t*-statistics of fixed effects from LMMs fit to the FPP data. *Freq* = word frequency; *Prime* = orthographic prime effect; *Voc* = vocabulary

As in the SPP analysis, the power transforms also affected the estimation of random-effect correlation patterns observed in the raw scale. Whereas larger word frequency and orthographic priming effects are associated with subjects who have higher mean RT in the raw scale ($r = -0.49$ and $r = -0.22$ respectively), the opposite relations appear in the inverse square-root scale ($r = 0.11$ and $r = 0.31$ respectively; Table 7).

As in Study 1, I performed the subsequent simulations to examine how the observed patterns in the analysis of the FPP might change as a function of the data-generating process and of subject and item sample sizes.

	Raw LMM			Inverse Square-Root LMM		
Fixed Effects	Estimate ($\hat{\gamma}$)	Std. Error	*t*	Estimate ($\hat{\gamma}$)	Std. Error	*t*
Intercept	666.44	3.33	199.99	-39.3509	0.0947	-415.66
x_1 (Frequency)	-32.65	1.66	-19.69	-0.9718	0.0500	-19.46
x_2 (Prime Type)	-21.02	1.02	-20.67	-0.6683	0.0256	-26.14
z_1 (Vocabulary)	-12.19	3.29	-3.70	-0.3426	0.0918	-3.74
x_1x_2 (Freq x Prime)	3.87	0.83	4.67	0.0602	0.0233	2.59
x_1z_1 (Freq x Voc)	3.75	0.93	4.04	0.0846	0.0260	3.26
x_2z_1 (Prime x Voc)	1.58	1.12	1.42	0.0123	0.0280	0.44
$x_1x_2z_1$ (F x P x V)	0.29	0.91	0.32	0.0145	0.0257	0.57
Random Effects	SD		Correlations	SD		Correlations
Subject ($\widehat{\boldsymbol{T}}_{subj}$)						
Intercept	89.55	1		2.4927	1	
x_1	8.05	-0.49	1	0.2178	0.11	1
x_2	18.31	-0.22	0.15	0.3294	0.31	0.25
Item ($\hat{\tau}_{00(item)}$)						
Intercept	29.26			0.9018		
Residual ($\widehat{\sigma^2}$)	125.50			3.5606		

Table 7. Parameter estimates from LMMs fit to raw and inverse-square-root-transformed FPP data. These estimates were used as the input parameters for the simulations performed in Studies 2.2 and 2.3 respectively.

Study 2.2: Simulations based on raw model estimates

Generating process. Setting up the generating process was identical to Study 1.2, except that a Gamma distribution with shape parameter = 0.15 and scale parameter = 675 was used to generate trial-level residuals (see Table 7). An intercept parameter of 625 was added to the simulated values to approximate the intercept estimated by the raw model.

Dataset generation and analysis. The datasets were generated to simulate a typical masked form priming experiment. In this experiment, all subjects see the same targets, but half the targets are preceded by orthographically related primes. Prime type (x_2) was counterbalanced in two lists so that targets preceded by orthographically related primes in one list are preceded by baseline primes in the other list. Based on their distributions in the megastudy, word frequency (x_1) and vocabulary (z_1) values were generated from standard normal distributions and the data was sorted so that word frequencies are matched between the prime type conditions.

The same data generation and analysis procedures were performed under the same subject and item sample size conditions as in Study 1 (SPP).

Performance and comparison measures. The same performance and comparison measures were obtained as in Study 1.2.

Results: Study 2.2

Performance measures: Raw model.

***Bias and coverage*.** Figure 18 shows the average percentage bias of the raw model for each of the effects as a function of subject and item sample size. Averaging across all subject and item sample sizes, the raw model underestimated the x_1, z_1, x_1x_2, and x_1z_1 effects by 5.63%, 1.39%, 4.69% and 6.04%, whereas it slightly overestimated the x_2 effect by 0.32%. The average bias for the null effects x_2z_1 and $x_1x_2z_1$ were negligible. Lastly, other than volatile estimates in models fit to small sample sizes, changes in sample size do not appear to be related to how much bias is incurred by the raw model.

Figure 19 shows the coverage for each effect in the raw model as a function of subject and item sample size. The results were very similar to those in Study 1.2. The raw model produced conservative coverage for x_1: averaging across all subject and item sample sizes, only 91.9% of the generated 95%-confidence intervals contained the true value of x_1. Moreover, increasing both subject and item sample sizes appeared to lower the coverage for this effect. All other effects had coverages that approximated the nominal 95% value. For these effects, changes in sample size did not affect their coverage estimates.

***Power/Type I error*.** The contour plots in Figure 20 show changes in power/Type I error for the raw model as a function of subject and item sample sizes. As expected, power for all existing effects increased as subject and item sample sizes increased. As in Study 1 (SPP), some effects are estimated to be more powerful than others: for instance, x_1 already has massive power at above 0.90 with 50 subjects and 50 items, whereas x_2's estimated power is only between 0.60 and 0.70 with these sample sizes and z_1, x_1x_2, and x_1z_1's power estimates are only below 0.20. Lastly, violating the normality assumption did

not seem to increase Type I error rates for the null effects: regardless of subject and item sample size, Type I error estimates for x_2z_1 and $x_1x_2z_1$ remained below 7%.

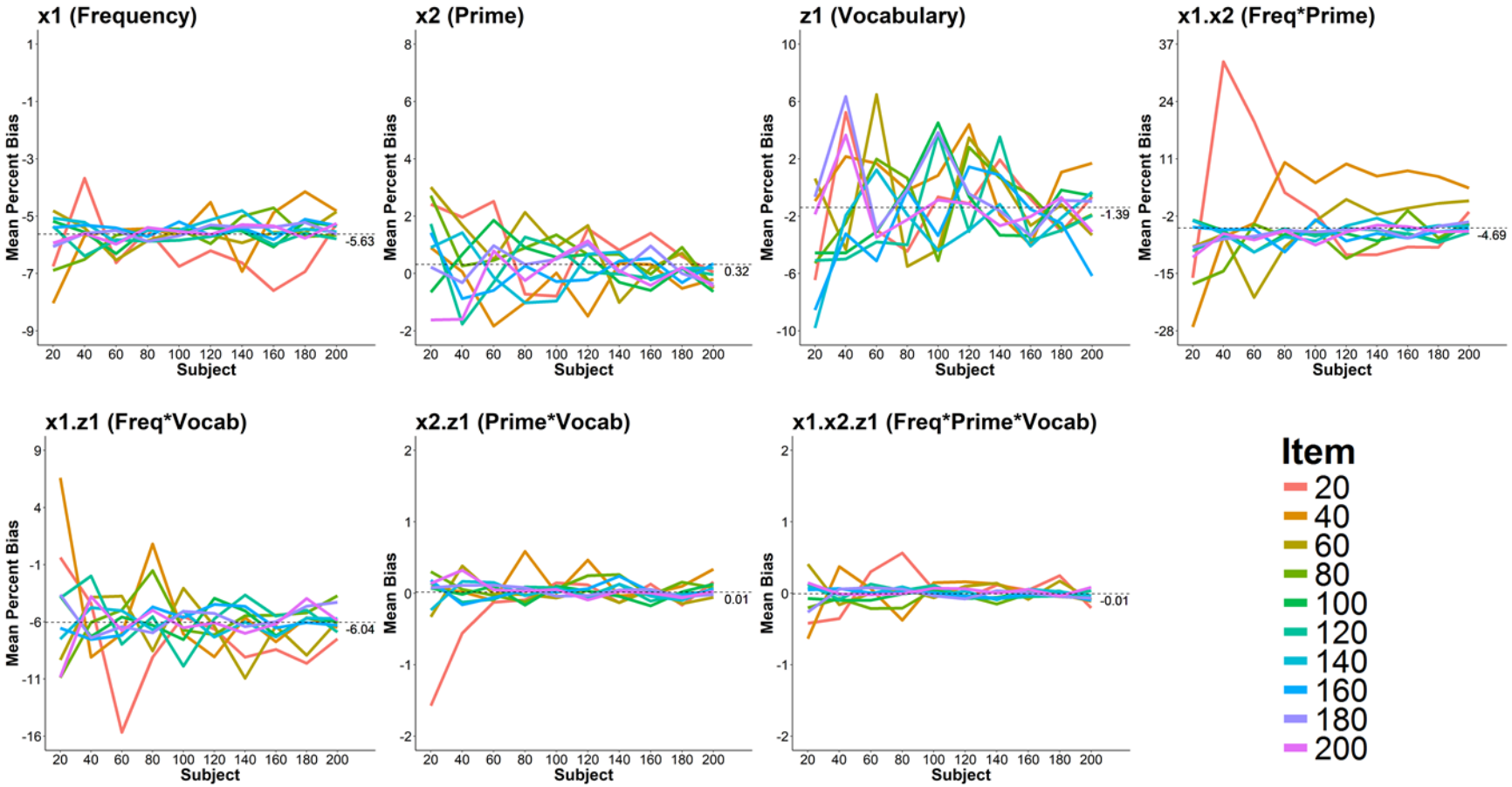

Figure 18. Average bias for each of the fixed-effect estimates in the raw model as a function of subject and item sample sizes. Simulation based on the model fit to the raw FPP data.

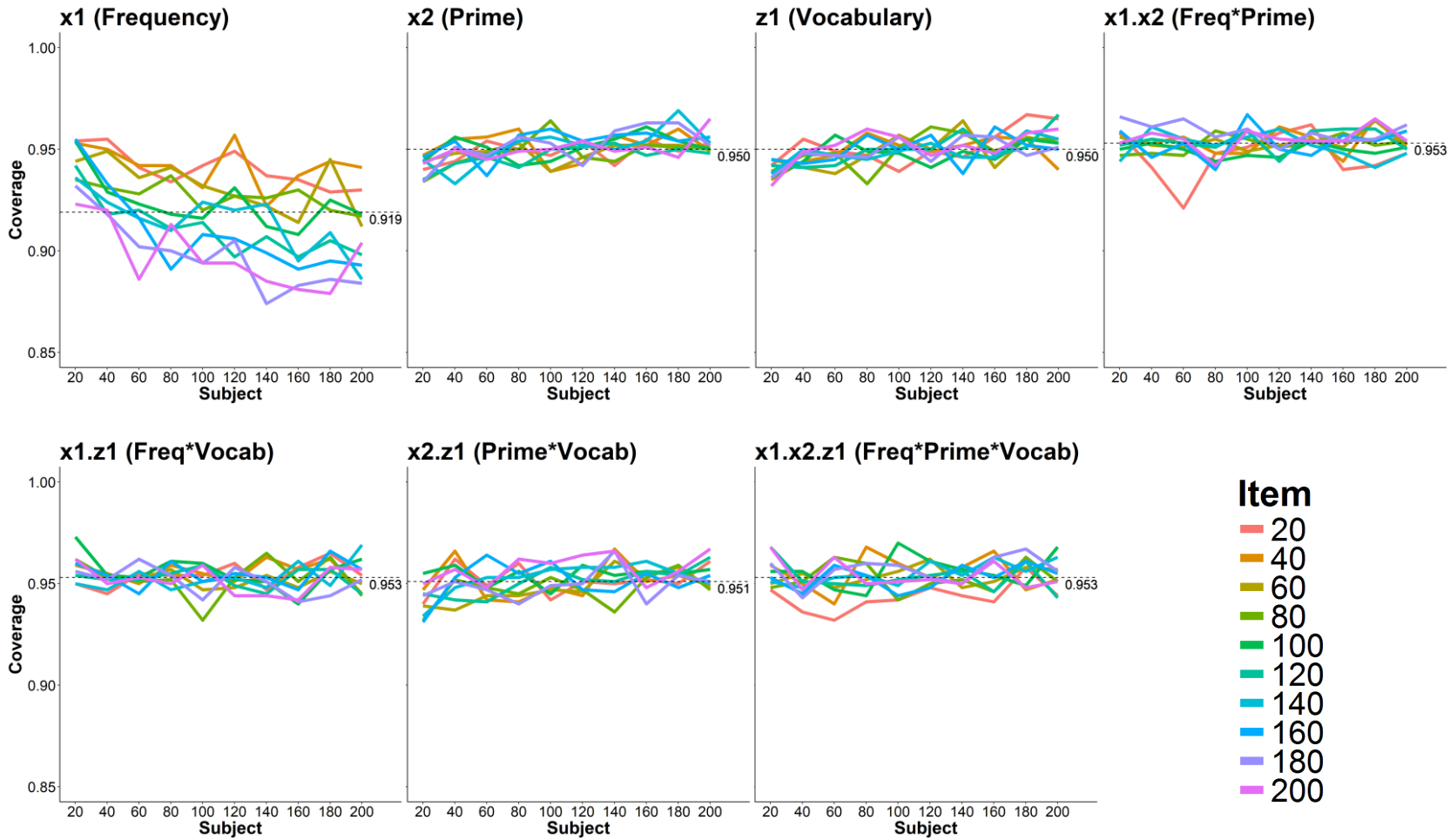

Figure 19. Coverage for each of the fixed effects in the raw model as a function of subject and item sample sizes. Simulation based on the model fit to the raw FPP data.

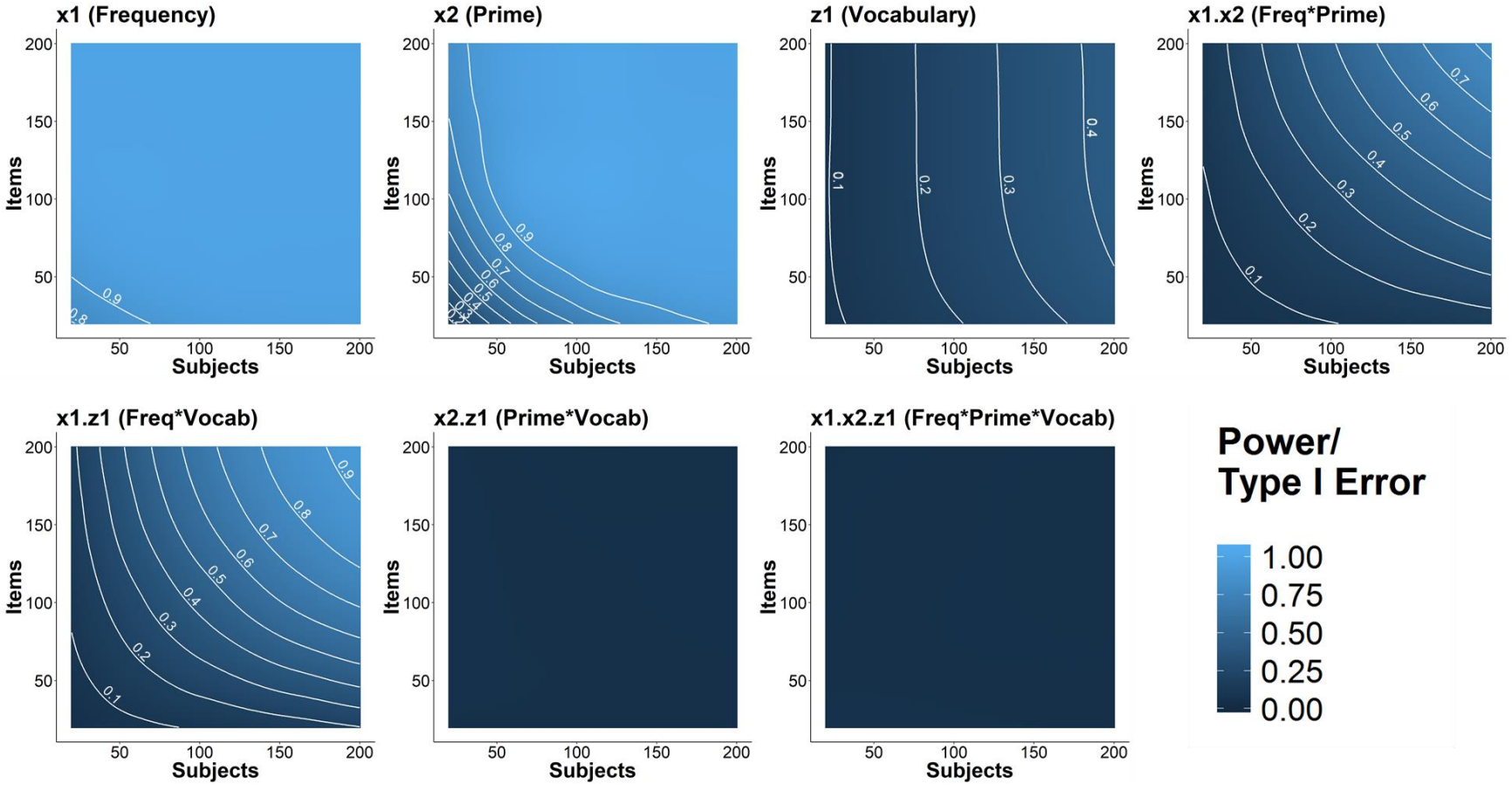

Figure 20. Power (Type I error) contour plots for the raw model as a function of subject and item sample size. The contour plots for $x_2 z_1$ and $x_1 x_2 z_1$ are Type I error contour plots; the rest are power contour plots. Simulation based on the model fit to the raw FPP data.

Comparison measures: Transformed vs. raw model. Figure 21 shows the proportion of datasets where both the raw and transformed model converged as a function of subject and item sample sizes. As in Study 1 (SPP), item sample sizes appear to have a greater effect on this convergence consistency, such that when the sample had fewer items, models that converged when fit to raw data tended not to converge when fit to transformed data.

Subsequent comparisons between the raw and the transformed models were then made on datasets where both the raw and transformed model converged.

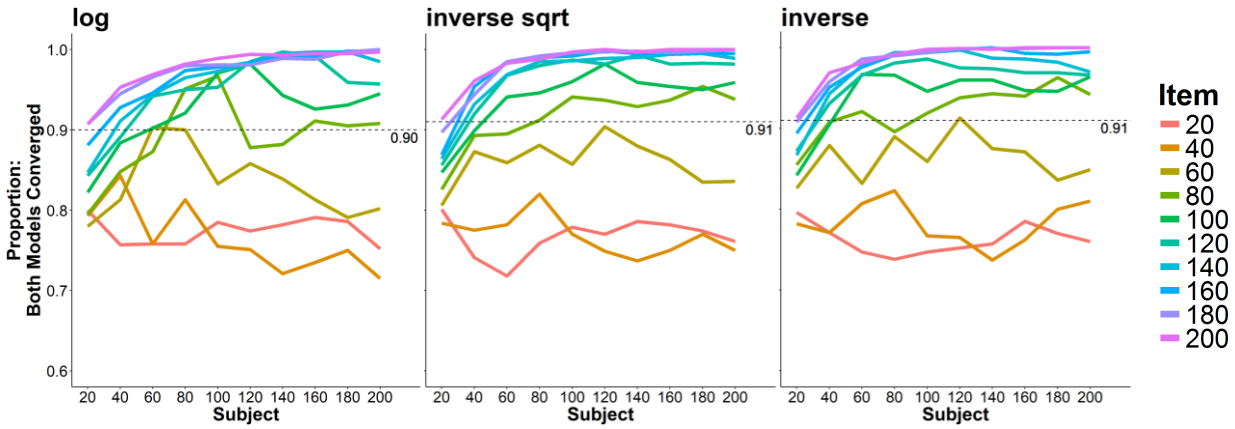

Figure 21. Proportion of datasets where the transformed models converged given that the raw model also converged, expressed as a function of subject and item sample sizes. Simulation based on the model fit to the raw FPP data.

Log model. The log model converged in 90,100 of the 100,000 datasets where the raw model converged. As in the raw model, increasing subject and item sample sizes also increased power in the log model for all the existing effects as shown in Figure 22. Moreover, Type I error rates for the null effects still remained below 7% regardless of sample size.

The contour plots in Figure 23 show the estimated difference in power/Type I error between the raw and the log model for each of the effects as a function of subject and item sample sizes. For x_1 and x_2, the log model was estimated to have as much as 8% and 14% more power than the raw model given smaller sample sizes respectively. For the $x_1 z_1$ interaction, the log model was estimated to have as much

as 6% more power as sample sizes increased. Differences in power for z_1 and x_1x_2 and Type 1 error for x_2z_1 and $x_1x_2z_1$ remained below 5% across all sample sizes.

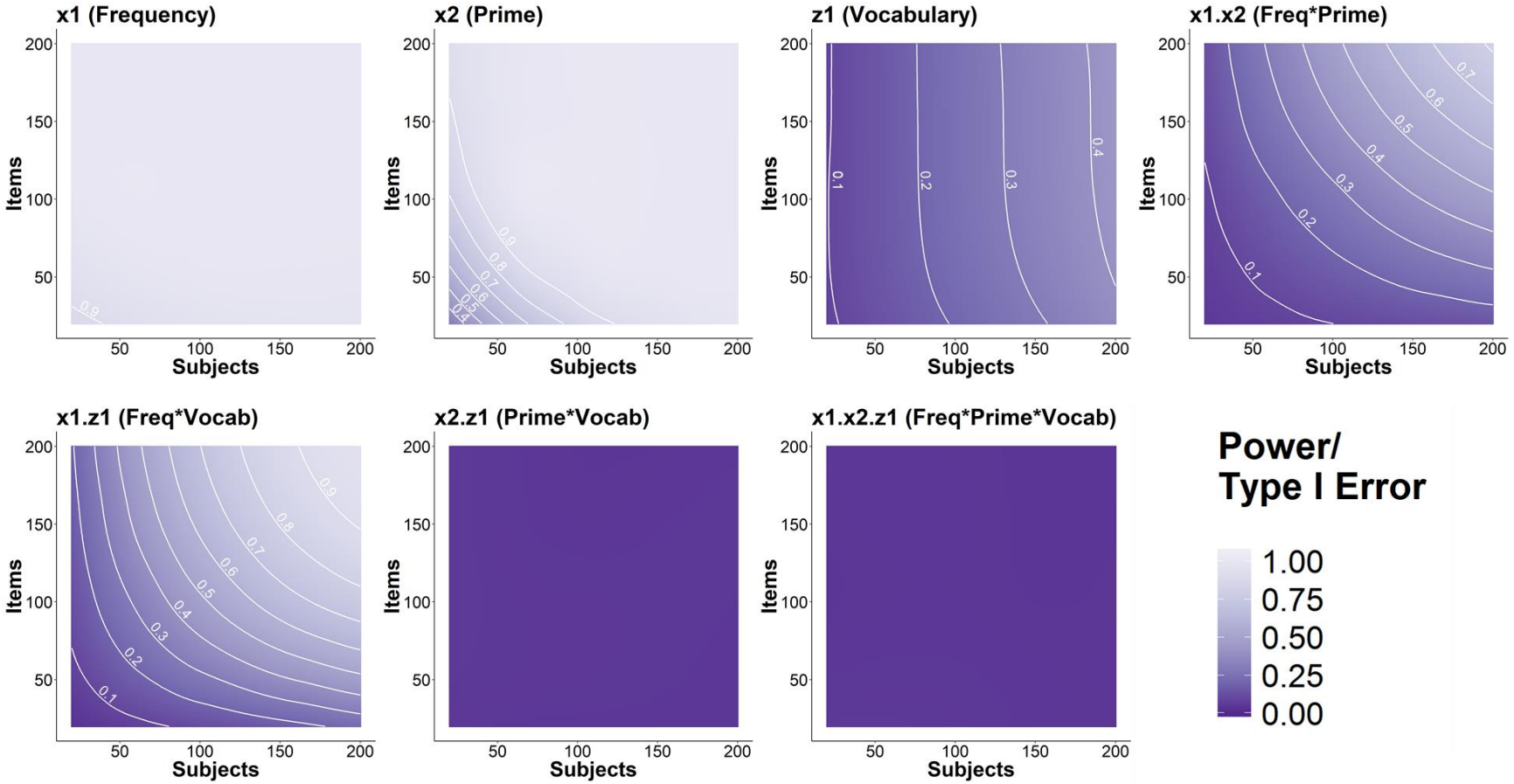

Figure 22. Power (Type I error) contour plots for the log model as a function of subject and item sample size. The contour plots for $x_2 z_1$ and $x_1 x_2 z_1$ are Type I error contour plots; the rest are power contour plots. Simulation based on the model fit to the raw FPP data.

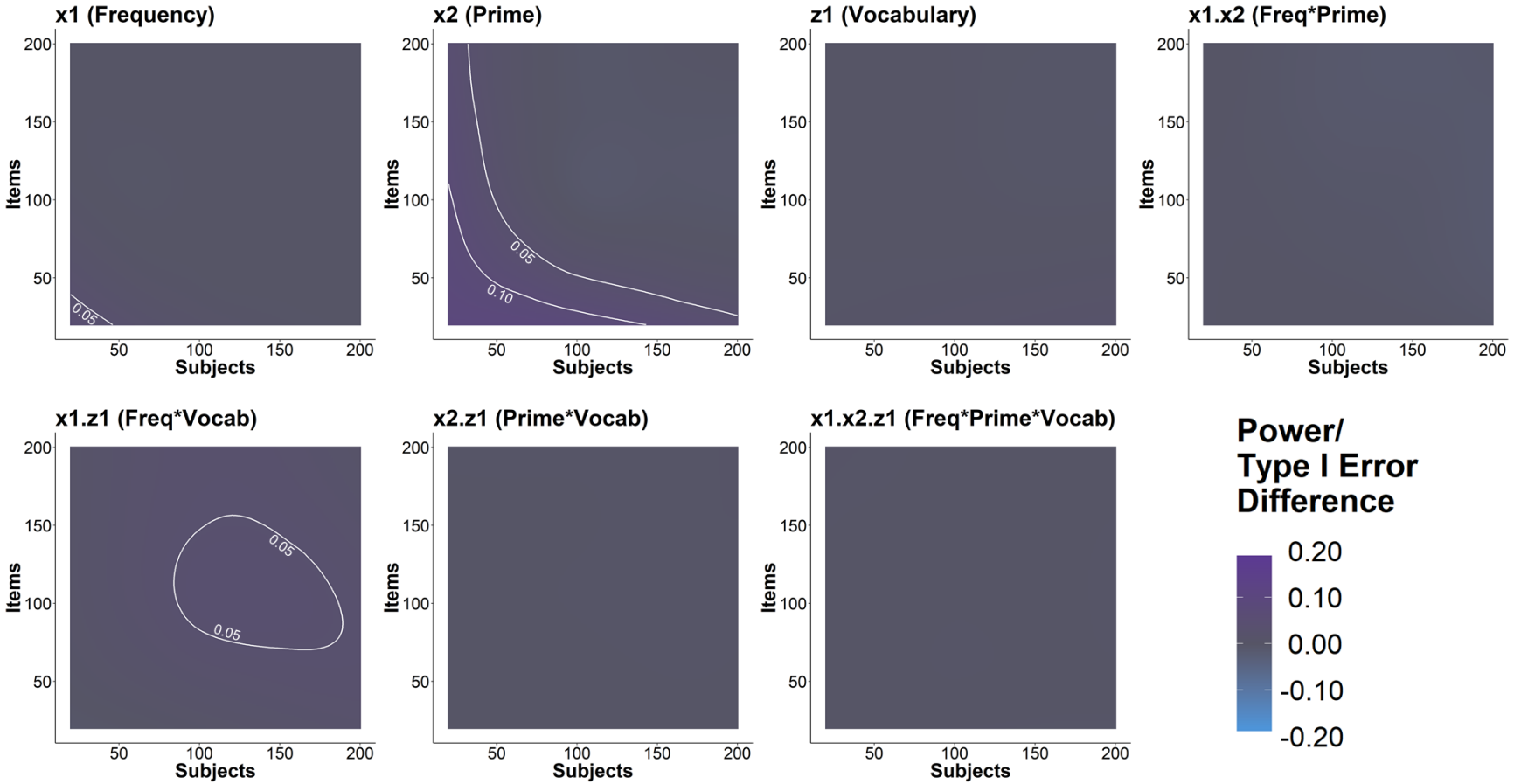

Figure 23. Power (Type I error) difference contour plots (log – raw) as a function of subject and item sample size. The contour plots for x_2z_1 and $x_1x_2z_1$ are a Type I error contour plots; the rest are power contour plots. Blue indicates that raw model has more power (Type I error) than the log model. Purple indicates that log model has more power (Type I error) than the raw model. Simulation based on the model fit to the raw FPP data.

Inverse square-root model. The inverse square-root model converged in 91,247 of the 100,000 datasets where the raw model converged. Figure 24 shows the general increase in power for all existing effects as sample size increased, whereas Type I error rates for the null effects remained below 8% regardless of sample size.

Figure 25 shows the estimated differences in power/Type I error between the raw and the inverse square-root model for each of the effects as a function of subject and item sample sizes. For x_1 and x_2, the inverse square-root model was estimated to have as much as 9% and 19% more power than the raw model given smaller sample sizes respectively. This indicates that the inverse square-root model has more power for these main effects than the log and raw models in smaller sample sizes. As for the interactions, the *raw* model notably had greater power in detecting the x_1x_2 effect, where it had as much as 10% more power than the than inverse square-root model as sample sizes increased. Differences in power for z_1 and x_1z_1 and in Type I error for x_2z_1 and $x_1x_2z_1$ between the raw and inverse square-root model remained below 5% across sample sizes. Recall that the optimal transform of the real RT data whose analysis was the basis for this simulation was closer in strength to the inverse square-root transform than the inverse transform.

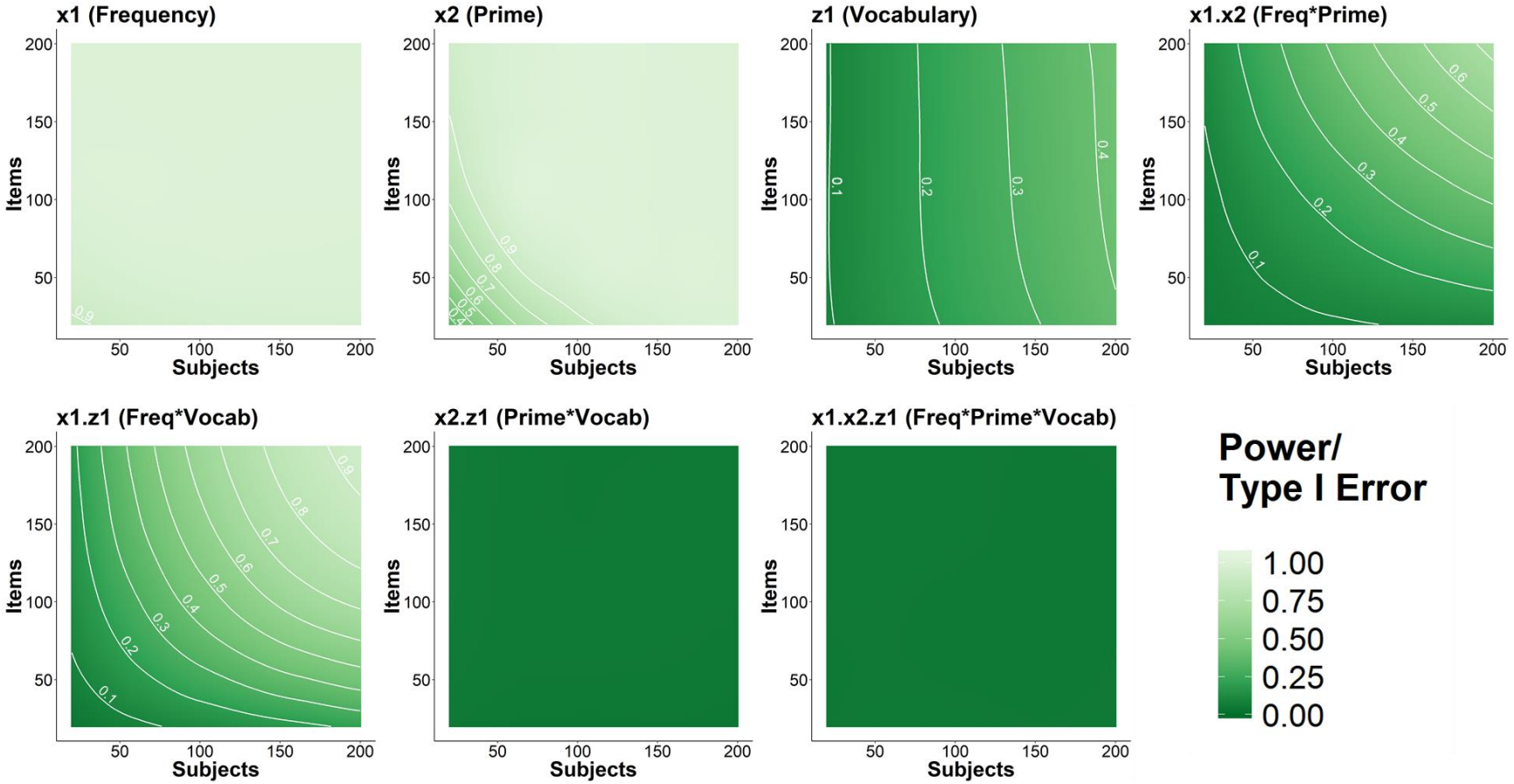

Figure 24. Power (Type I error) contour plots for the inverse square-root model as a function of subject and item sample size. The contour plots for x_2z_1 and $x_1x_2z_1$ are Type I error contour plots; the rest are power contour plots. Simulation based on the model fit to the raw FPP data.

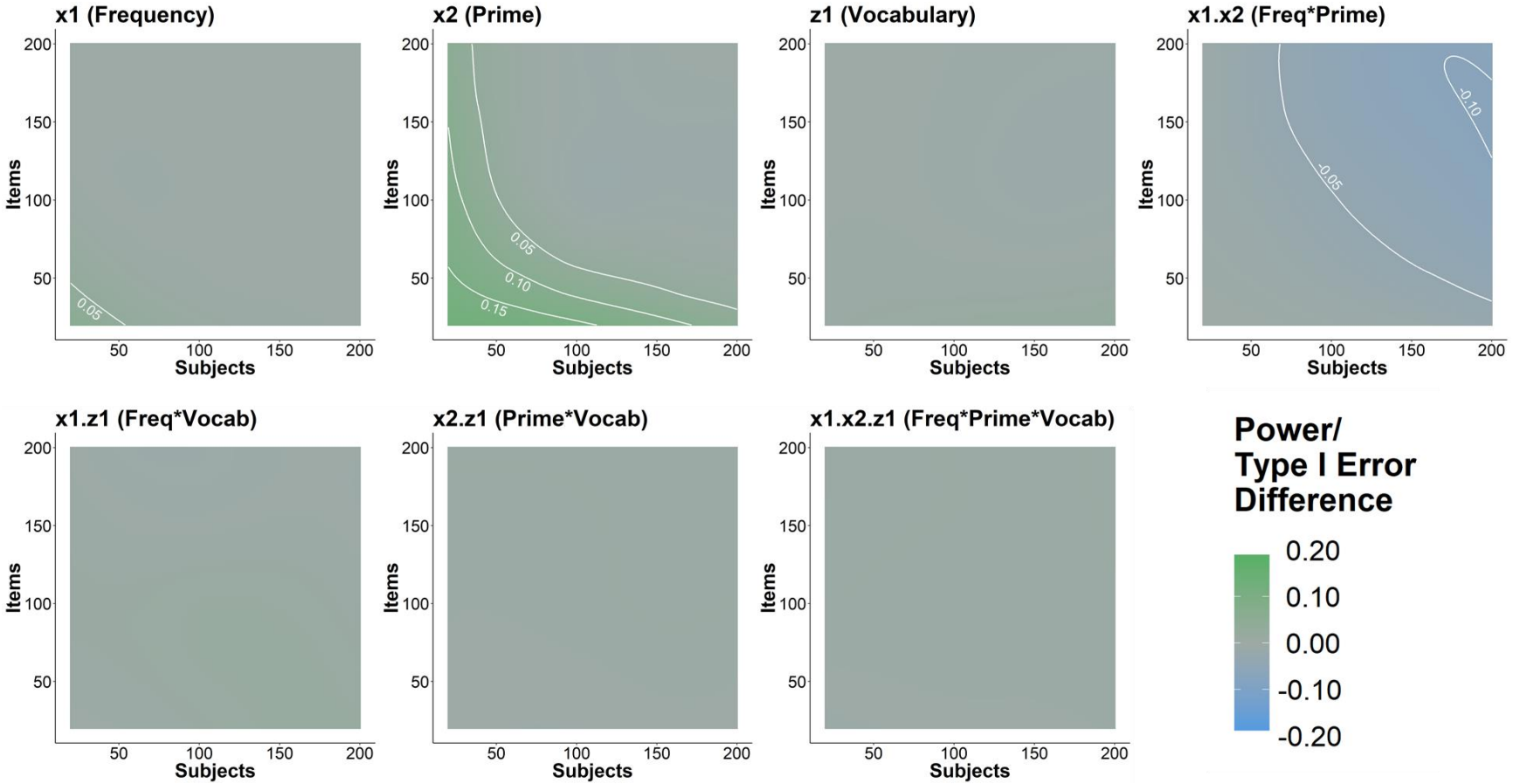

Figure 25. Power (Type I error) difference contour plots (inverse square-root – raw) as a function of subject and item sample size. The contour plots for x_2z_1 and $x_1x_2z_1$ are a Type I error contour plots; the rest are power contour plots. Blue indicates that raw model has more power (Type I error) than the inverse square-root model. Green indicates that inverse square-root model has more power (Type I error) than the raw model. Simulation based on the model fit to the raw FPP data.

Inverse model. The inverse model converged in 91,461 of the 100,000 datasets where the raw model converged. Figure 26 shows the general increase in power for all existing effects as sample size increased; Type I error rates for the null effects remained below 8% regardless of sample size.

Figure 27 shows the estimated differences in power/Type I error between the raw and the inverse model for each of the effects as a function of subject and item sample sizes. For x_1 and x_2, the inverse model was estimated to have as much as 10% and 24% more power than the raw model given smaller sample sizes respectively. This indicates that the inverse model is the most powerful model of the four LMMs in detecting the x_1 and x_2 effects given smaller sample sizes. As for the interactions, the *raw* model had greater power for the x_1x_2 and x_1z_1 effects, having as much as 25% and 12% more power as sample sizes increased for these effects than the inverse model respectively. Differences in power for z_1 and Type I error rates for x_2z_1 and $x_1x_2z_1$ between the raw and inverse model remained below 5% regardless of sample size.

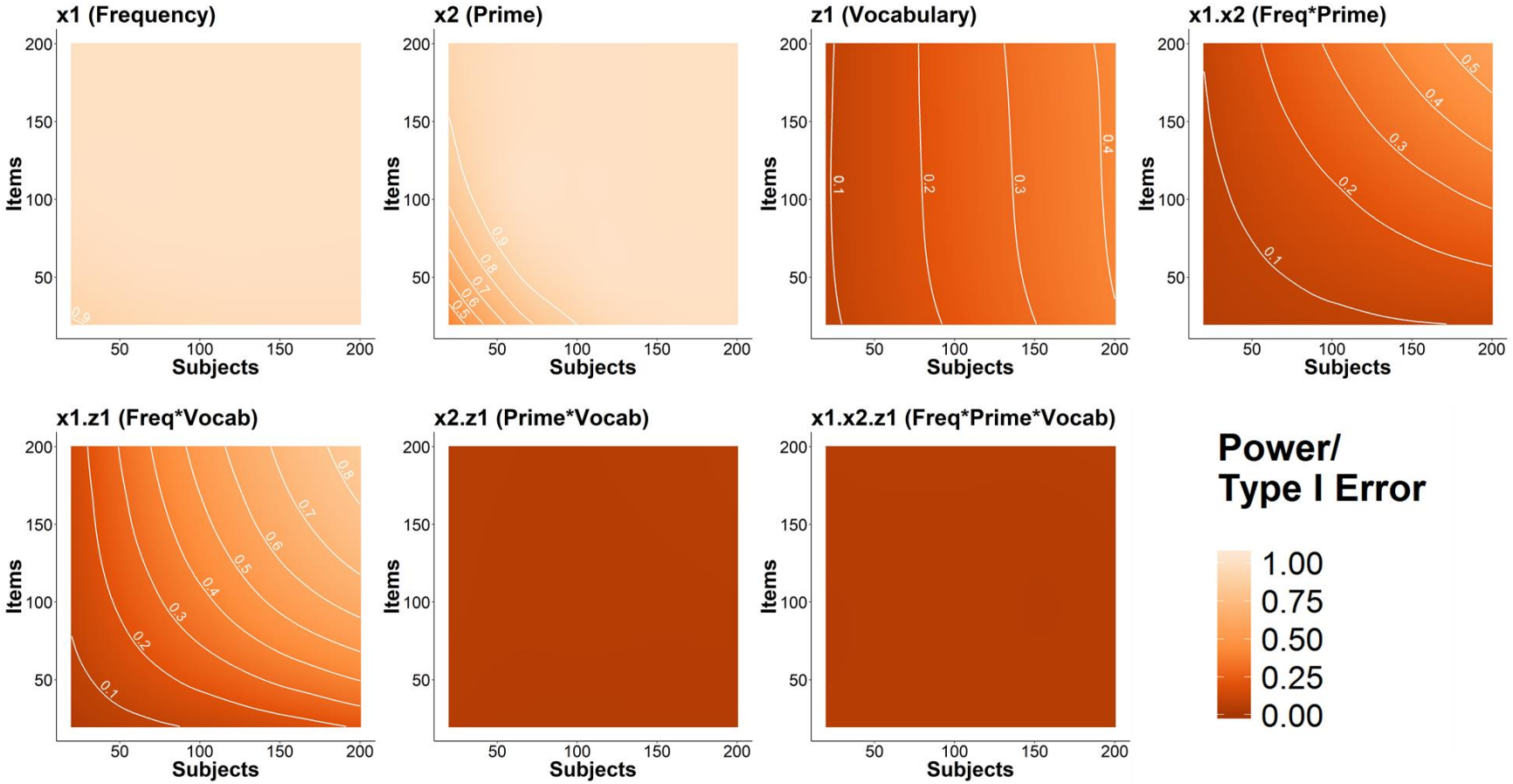

Figure 26. Power (Type I error) contour plots for the inverse model as a function of subject and item sample size. The contour plots for $x_2 z_1$ and $x_1 x_2 z_1$ are Type I error contour plots; the rest are power contour plots. Simulation based on the model fit to the raw FPP data.

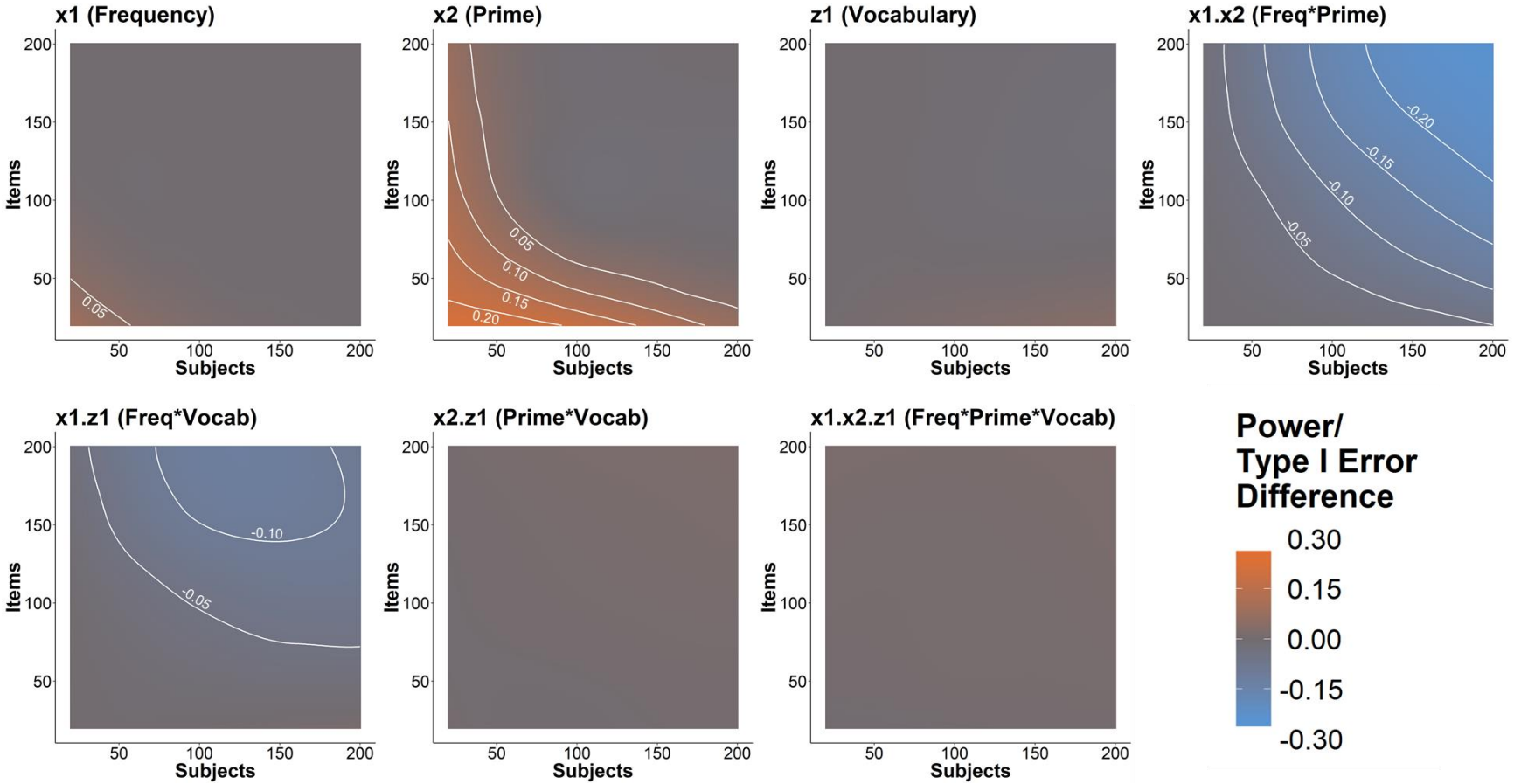

Figure 27. Power (Type I error) difference contour plots (inverse – raw) as a function of subject and item sample size. The contour plots for x_2z_1 and $x_1x_2z_1$ are a Type I error contour plots; the rest are power contour plots. Blue indicates that raw model has more power (Type I error) than the inverse square-root model. Orange indicates that inverse model has more power (Type I error) than the raw model. Simulation based on the model fit to the raw FPP data.

Summary. There are three main findings from this simulation, First, in violating the normality assumption, the raw model underestimated the x_1 and z_1 main effects and both the existing interaction effects. Second, across all RT scales, increasing sample sizes increased power for existing effects, but it did not affect Type I error rates for null effects. These results were practically identical to those obtained from Study 1.2.

Lastly and most notably, as the power transform became stronger, the transformed model continued to outperform the raw model in detecting the x_1 and x_2 main effects in smaller sample sizes, but it became *worse* than the raw model in detecting the existing interactions. This was despite the raw model *underestimating* both the main effects and the interaction effects. Differences in Type I error rates for the null interactions between the raw and transformed models remained negligible across sample sizes.

To address the possibility that these results were dependent on the generating process being based on the raw model, the simulation was performed again using a transformed model as the basis of the generating process. Because the optimal transform of the FPP data was closer in strength to inverse square-root transform than the inverse transform, the following simulation was based on the inverse square-root model.

Study 2.3: Simulations based on inverse square-root model estimates

Generating process. The parameters used in the generating process of the simulation are the significant fixed-effect estimates (implying 0 for x_2z_1 and $x_1x_2z_1$) and the random-effect estimates obtained from the inverse square-root model (summarized in Table 7). With the Box-Cox procedure revealing that the optimal transform is closer in strength to the inverse square-root transform, a normal distribution with mean = 0 and variance = 3.5606^2 = 12.6778 (which are the estimates obtained from the inverse square-root model) was used to generate trial-level residuals.

Dataset generation and analysis. This was identical to Study 2.2, except that the simulated inverse square-root RTs were back-transformed only into raw RTs to compare the performance of the models fit to these two RT scales.

Comparison measures: Inverse square-root vs. raw model. The same performance measures were obtained as in Study 1.2.

Results: Study 2.3

Figure 28 shows the proportion of datasets for which both the raw and inverse square-root models converged as a function of subject and item sample sizes. As in the prior simulations, models that converged when fit to the raw data did not necessarily converge when fit to the inverse-square-root-transformed data. Smaller item sample sizes seemed to exacerbate this issue. Subsequent results were evaluated on the 92,291 datasets where both the inverse square-root model and the raw model converged.

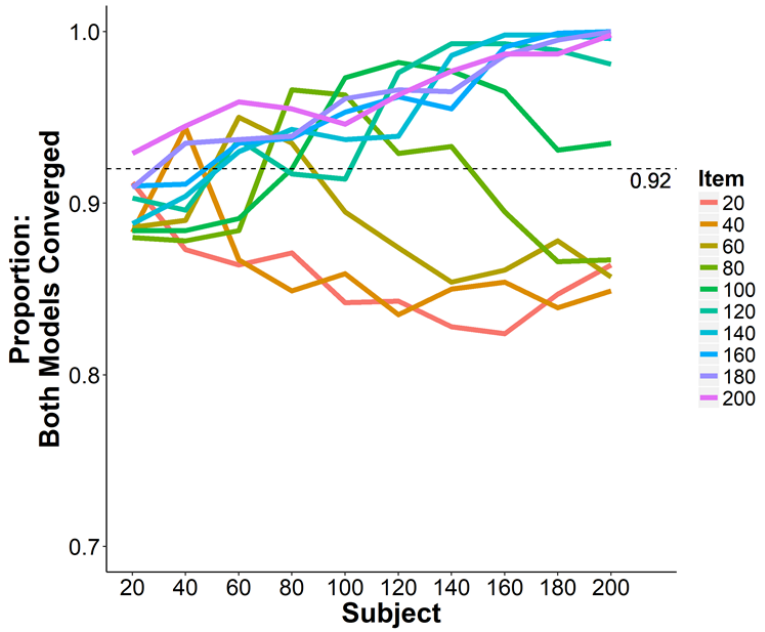

Figure 28. Proportion of datasets where the inverse square-root models converged when the raw model also converged, expressed as a function of subject and item sample sizes. Simulations based on the model fit to inverse-square-root-transformed FPP data.

The power contour plots for the raw model and inverse square-root model in Figures 29 and 30 show the general increase in power for the existing effects as subject and item sample sizes increase. Type

I error for the x_2z_1 and $x_1x_2z_1$ effects remained below 7% for both the raw and inverse square-root model.

Differences between the raw and inverse square-root models' power estimates are seen in the power difference contour plots in Figure 31. Notably, the raw model's power for the x_1x_2 and x_1z_1 interactions became as much as 41% and 16% higher than the inverse square-root model's respectively as sample sizes increased. There was no systematic power or Type I error advantage observed with the inverse square-root model as sample sizes increased for any of the other fixed effects. In fact, differences in power and Type I error estimates between the raw and inverse square-root transformed models for all other fixed effects remained 8% or lower across all sample size conditions. Ultimately, the raw model outperformed the inverse-square root model despite the fact that the simulation's generating process was based on the model fit to the inverse-square-root-transformed FPP data.

Overall, the results look very similar to those observed in Study 2.2, thereby providing stronger evidence than Study 1 in showing that the choice of underlying generating process does not change the pattern of results obtained from the simulations.

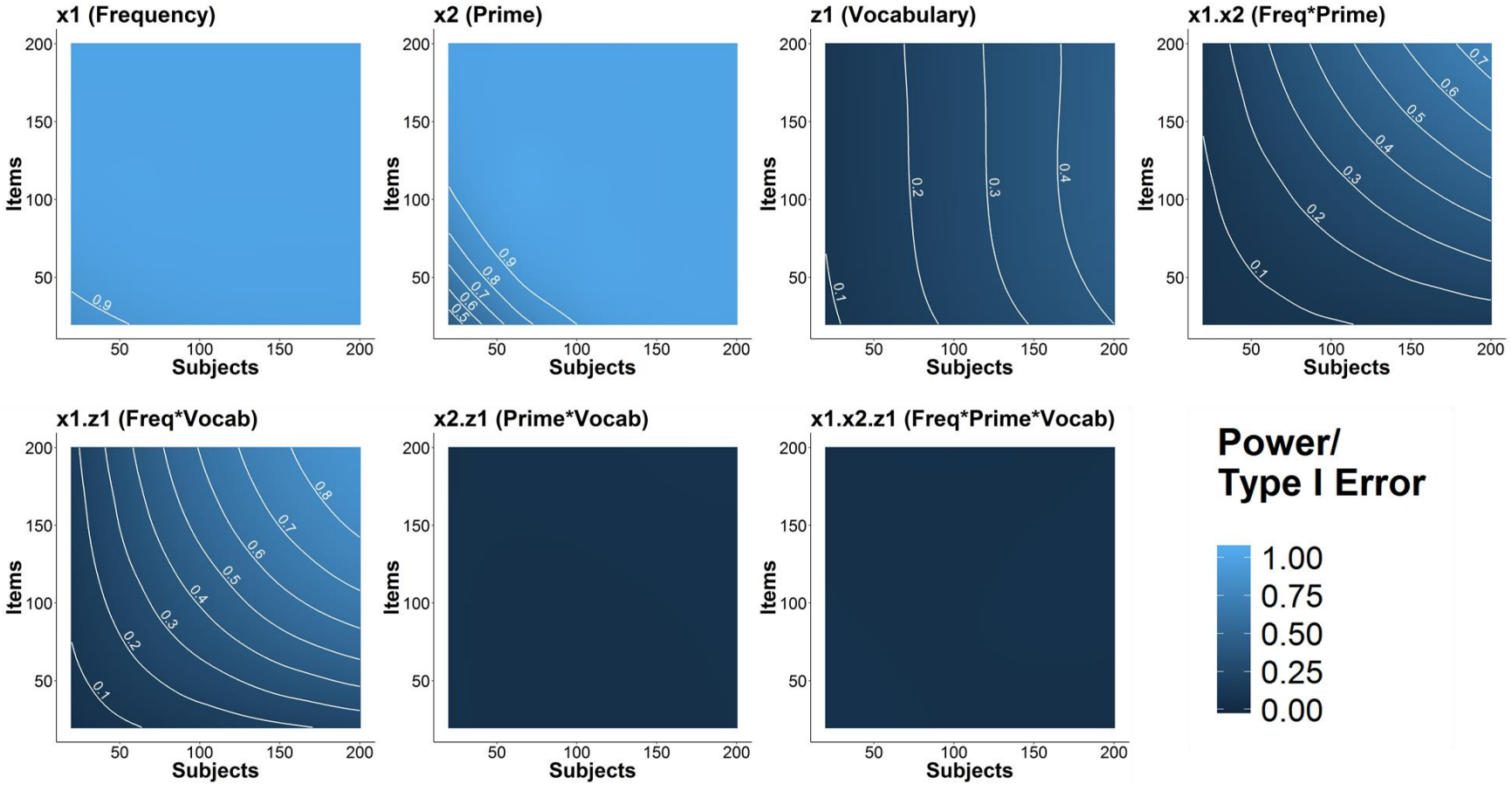

Figure 29. Power (Type I error) contour plots for the raw model as a function of subject and item sample size. The contour plots for $x_2 z_1$ and $x_1 x_2 z_1$ are Type I error contour plots; the rest are power contour plots. Simulation based on the model fit to the inverse-square-root-transformed FPP data.

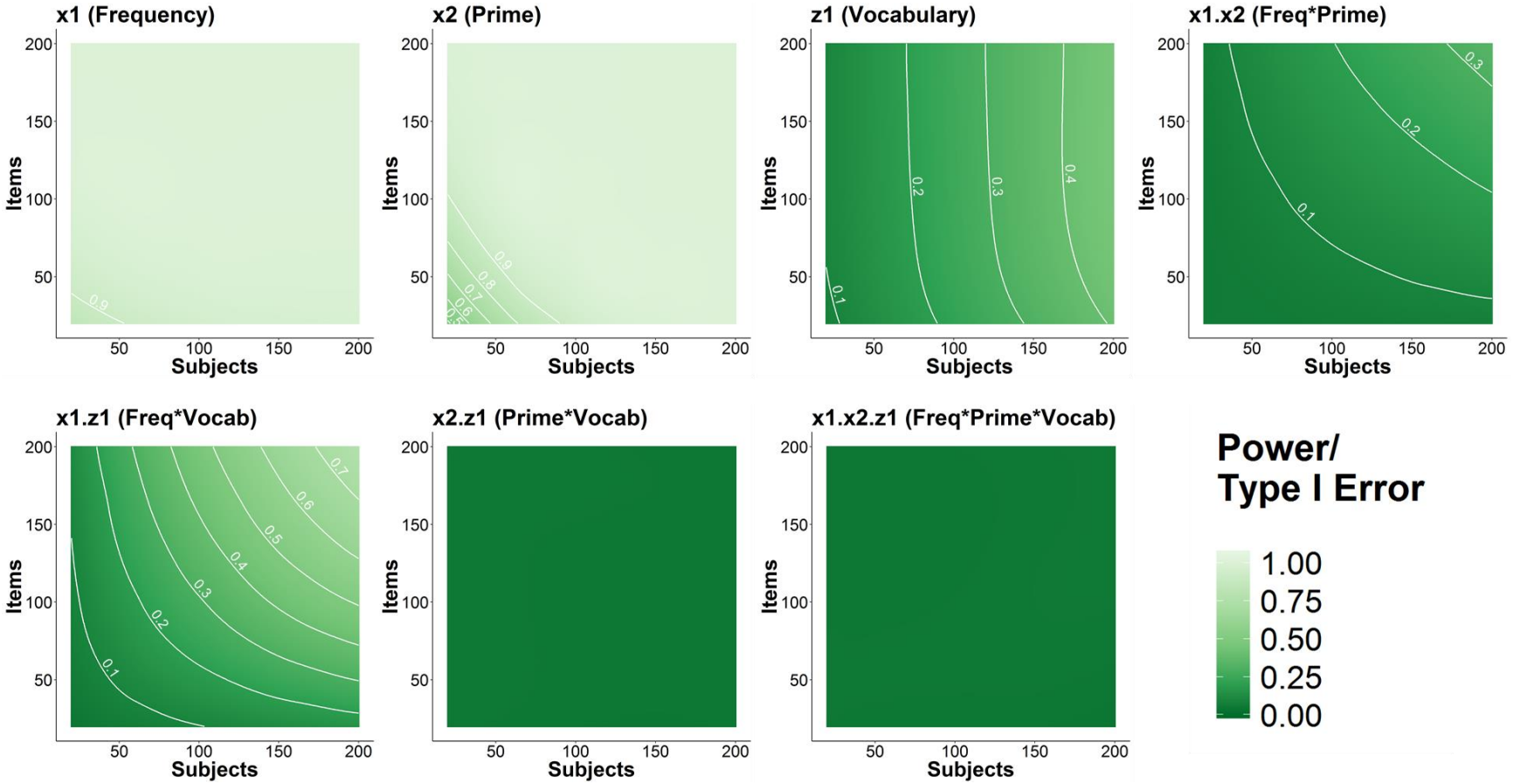

Figure 30. Power (Type I error) contour plots for the inverse square-root model as a function of subject and item sample size. The contour plots for x_2z_1 and $x_1x_2z_1$ are Type I error contour plots; the rest are power contour plots. Simulation based on the model fit to the inverse-square-root-transformed FPP data.

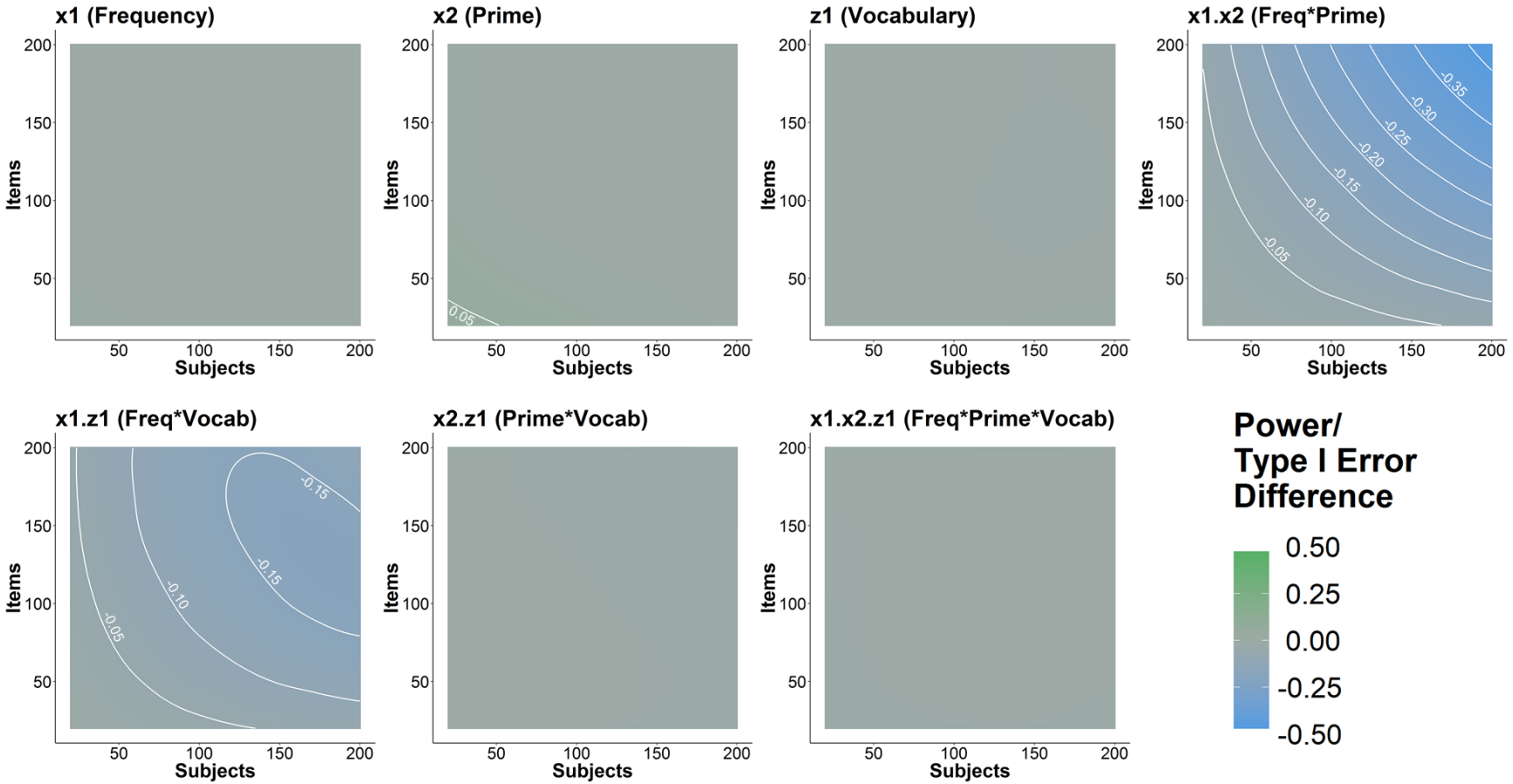

Figure 31. Power (Type I error) difference contour plots (inverse square-root – raw) as a function of subject and item sample size. The contour plots for x_2z_1 and $x_1x_2z_1$ are a Type I error contour plots; the rest are power contour plots. Blue indicates that raw model has more power (Type I error) than the inverse square-root model. Green indicates that inverse square-root model has more power (Type I error) than the raw model. Simulation based on the model fit to the inverse-square-root-transformed FPP data.

Discussion

The FPP analysis revealed qualitatively similar results to those obtained from the SPP analysis: as stronger power transforms were applied to the data, the *t*-statistics of the main effects were only slightly affected but those of the two-way interactions systematically decreased. As in the SPP, power transforms also altered random effect correlation patterns present in the raw scale.

However, unlike the results in Study 1, the FPP LMM results were supported by the simulations based on both the raw and inverse square-root LMMs as generating processes. The transformed LMMs were more powerful than the raw LMM in detecting the main effects of x_1 and x_2 in small sample sizes, which is again consistent with the ANOVA simulations on small sample sizes (Levine & Dunlap, 1982; Ratcliff, 1993). But analogous to the systematic decrease in the *t*-statistics of the two-way interactions as stronger power transforms were applied to the FPP data, the transformed models became less sensitive in detecting x_1x_2, and x_1z_1 as sample sizes increased and as stronger power transforms were applied to the simulations.

Interestingly, the power estimates I obtained for the orthographic priming effect (x_2) were similar to those obtained by Brysbaert & Stevens (2018) in their FPP analysis and simulations. This was despite their different way of dichotomizing the 28 priming conditions in the FPP and the much simpler LMM they fit to the FPP (i.e., the only fixed effect was x_2). This correspondence in results provides converging evidence about the reliability of the current simulations.

Study 3: English Lexicon Project

The ELP (Balota et al., 2007) is the most well-known word-recognition megastudy. It is currently the largest descriptive database of word recognition, with 1,290 subjects responding to a subset of 40,481 words. I analyzed the lexical decision subset of the ELP which consists of 816 participants, each of whom responded to a ~1,700-word subset of all the words tested in the ELP. Consequently, ~34 observations were made on each of the 40,481 words available in the ELP. More details about the procedure are available in Balota et al. (2007).

In Study 3, I focused on the following three predictors in the ELP and their interactions:

1) The target word frequency (x_1), which is a continuous word-level characteristic representing the target's $\log_{10}$ frequency per million words based from the HAL corpus, a corpus of word frequencies from Usenet newsgroups during February 1995 (Lund & Burgess, 1996);
2) The target word length (x_2), which is a continuous word-level characteristic indicating the number of letters in a word; and
3) Vocabulary (z_1), which is a continuous person-level characteristic representing the number of items a subject got correct on a variant of the Shipley vocabulary test

Study 3.1: Analyzing ELP Data

Data preprocessing. Only RTs from accurate responses to lexical targets (i.e., word strings correctly identified as words) were analyzed. Prior to screening the data for outliers, word frequency and word length were item grand-mean z-transformed, and vocabulary was person grand-mean z-transformed. The same preprocessing and model fitting procedures as in Studies 1 and 2 were applied to the data, except that outlying observations were identified simply as those which are beyond +/- 2.5 SDs of the subject's mean instead of using the Van Selst & Jolicoeur (1994) outlier criterion procedure. Unlike Studies 1 and 2, there are no experimental conditions by whose sample size the outlier criterion needs to be adjusted in the ELP, eliminating the need to use the Van Selst & Jolicoeur procedure. This preprocessing procedure resulted in 4.86% to 10.41% of observations being dropped from further analyses across RT scales. Twenty-six subjects were further excluded due to missing vocabulary scores, resulting in 727 to 758 subjects being retained for analyses.

Fitting the LMMs and identifying the optimal power transform. The main effects of word frequency, word length, and vocabulary and all interaction effects were entered into cross-classified LMMs, one for each RT scale. I specified random intercepts for subjects and items and random subject slopes for the word frequency (x_1) and word length (x_2) effects.

After the models were fit, the optimal power transform for the preprocessed data was identified using the Box-Cox procedure, which revealed that the data is optimally normalized using λ = -0.46,

which roughly corresponds to the inverse square-root transform. However, the residual QQ plots in Figure 32 show that while the stronger power transforms reduced the residuals' positive skew more, they also introduced more negative skew. Thus, the Box-Cox procedure identified the inverse square-root transform as the optimally normalizing transform presumably because it minimized this trade-off.

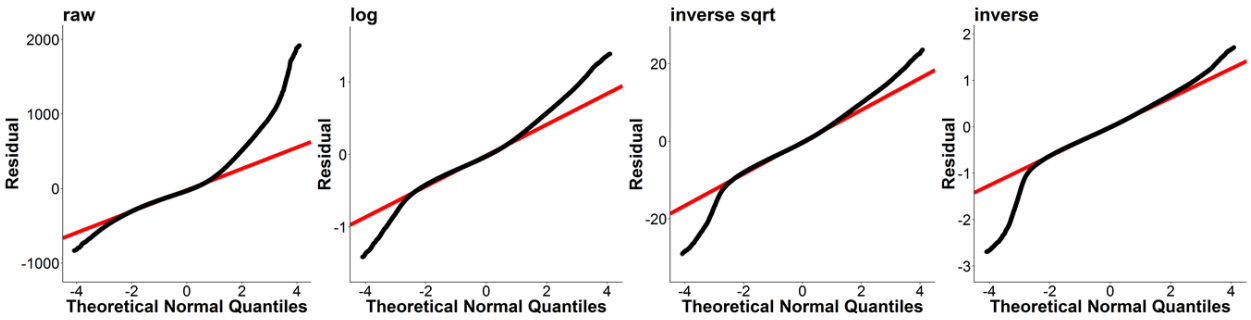

Figure 32. Trial-level residual QQ plots of models fit to ELP data

The *t*-statistics across all four models show a more extreme but similar pattern of changes to those observed in both the SPP and FPP analyses. Table 8 shows that the *t*-statistics for x_1 and x_2 increased as a stronger power transform was applied to the data. However, all two-way interaction *t*-statistics systematically shrank as the transform became stronger, so much so that the *t*-statistic for the x_1z_1 effect no longer meets the critical value of $|2|$ and is thus nonsignificant in the inverse model. Remarkably, the positive three-way interaction estimate in the raw model changed sign in all of the transformed models and became negative. That is, while the frequency-length interaction gets weaker as vocabulary increases in the raw model, the same interaction gets stronger as vocabulary increases in all of the transformed models. Such change in an effect's direction is new in the ELP analysis.

RT Scale	x_1 (Freq)	x_2 (Len)	z_1 (Voc)	x_1x_2 (Freq x Len)	x_1z_1 (Freq x Voc)	x_2z_1 (Len x Voc)	$x_1x_2z_1$ (F x L x V)
Raw	-73.49	45.19	-7.65	-22.38	6.89	-6.01	3.51
Log	-89.17	51.06	-8.21	-11.61	5.39	-5.44	**-1.50**
Inverse Sqrt	-94.47	54.39	-8.10	-7.45	3.55	-4.87	**-3.09**
Inverse	-95.97	57.14	-8.09	-3.61	**1.12**	-3.69	**-5.11**

Table 8. *t*-statistics of fixed effects from LMMs fit to the ELP data. Effects that changed significance and direction from raw model are boldfaced. *Freq* = word frequency; *Len* = word length; *Voc* = vocabulary

Beyond the fixed effects, the power transforms only altered the magnitude of the random-effect correlation patterns and not the direction of the patterns, unlike the patterns observed in the SPP and the FPP. Whereas the size of the word frequency and word length effects for each subject is strongly related to subjects' average RT in the raw scale ($r = -0.73$ and $r = 0.68$ respectively), these correlations manifest as only modest in the inverse square-root scale ($r = -0.33$ and $r = 0.43$ respectively; Table 9).

Model	Raw LMM			Inverse Square-Root LMM		
Fixed Effects	Estimate ($\hat{y}$)	Std. Error	*t*	Estimate ($\hat{y}$)	Std. Error	*t*
Intercept	737.30	4.03	183.00	-37.7298	0.1040	-362.75
x_1 (Frequency)	-57.78	0.79	-73.49	-1.4983	0.0156	-94.47
x_2 (Length)	43.89	0.97	45.19	1.0928	0.0201	54.39
z_1 (Vocabulary)	-31.58	4.13	-7.65	-0.8560	0.1056	-8.10
x_1x_2 (Freq x Len)	-7.91	0.35	-22.38	-0.0639	0.0086	-7.45
x_1z_1 (Freq x Voc)	5.02	0.73	6.89	0.0484	0.1366	3.55
x_2z_1 (Len x Voc)	-5.63	0.94	-6.01	-0.0906	0.0186	-4.87
$x_1x_2z_1$ (F x L x V)	0.70	0.20	3.51	-0.0137	0.0044	-3.09
Random Effects	SD	Correlations		SD	Correlations	
Subject						
Intercept	107.78	1		2.8284	1	
x_1	18.12	-0.73	1	0.3393	-0.33	1
x_2	23.78	0.68	-0.59	0.4798	0.43	-0.29
Item						
Intercept	62.74			1.5745		
Residual	191.60			4.4688		

Table 9. Parameter estimates from LMMs fit to raw and inverse-square-root-transformed ELP data. These estimates were used as the input parameters for the simulations performed in Studies 3.2 and 3.3 respectively.

As in Studies 1 and 2, I performed the subsequent simulations to examine how different data-generating processes and subject and item sample sizes might influence the observed patterns in the analysis of the ELP.

Study 3.2: Simulations based on raw model estimates

Generating process. Setting up the generating process was identical to Study 1.2, except that a Gamma distribution with shape parameter = 0.1 and scale parameter = 3000 was used to generate trial-level residuals (see Table 9). An intercept parameter of 650 was added to the simulated values to approximate the intercept estimated by the raw RT LMM.

Dataset generation and analysis. The datasets were generated to simulate a smaller version of the lexical decision component of the ELP. In this study, lexical decision data are obtained on specific word stimuli, and all subjects respond to the same words. Because there is no manipulation in this study, no counterbalancing was performed in the simulations. Instead, because word frequency (x_1) is -0.35

correlated with word length (x_2) in the ELP, word frequency and length values were generated from a standard bivariate normal distribution with a -0.35 correlation. Lastly, vocabulary (z_1) scores were generated from a standard normal distribution based on its distributional characteristics in the megastudy.

The same data generation procedures were performed under the same subject and item sample size conditions as in Study 1. All datasets were finally analyzed using the same preprocessing and fitting procedure described in Study 3.1.

Performance and comparison measures. The same performance and comparison measures were obtained as in Study 1.2.

Results: Study 3.2

Performance measures: raw model.

***Bias and coverage*.** Figure 33 shows the average percentage bias of the raw model for each of the effects as a function of subject and item sample size. Averaging across all subject and item sample sizes, the raw model underestimated all the effects in the model: the x_1, x_2, and z_1 main effects by 7.15%, 7.46%, and 9.88% respectively; the x_1x_2, x_1z_1, and x_2z_1 two-way interaction effects by 8.73%, 11.70%, and 12.15% respectively; and the $x_1x_2z_1$ three-way interaction by 10.08%. These results are consistent with those observed in Studies 1 and 2. Moreover, changes in sample size did not appear to be related to how much bias is incurred by the raw model, other than for the z_1 effect where the raw model incurred greater bias in samples with fewer items.

Figure 34 shows the coverage for each effect in the raw model as a function of subject and item sample size. Averaging across all subject and item sample sizes, the coverage for x_1 and x_2 were conservative at 91.3% and 92.8% respectively, and increasing both subject and item sample sizes appear to lower the coverage for these effects. On the other hand, the coverage for z_1 and all the interaction effects approximated the nominal 95% value and were less affected by changes in sample size.

***Power*.** The contour plots in Figure 35 show changes in power for the raw model as a function of subject and item sample sizes. As expected, power for all the effects increased as subject and item sample

sizes increased. As in Studies 1 and 2, some effects are estimated to be more powerful than others: for instance, x_1 already has massive power at above 0.90 with 50 subjects and 50 items, whereas x_2's estimated power is only between 0.80 and 0.90 with these sample sizes and z_1's is only between 0.30 and 0.40. Power for all interaction effects with these sample sizes are below 0.20.

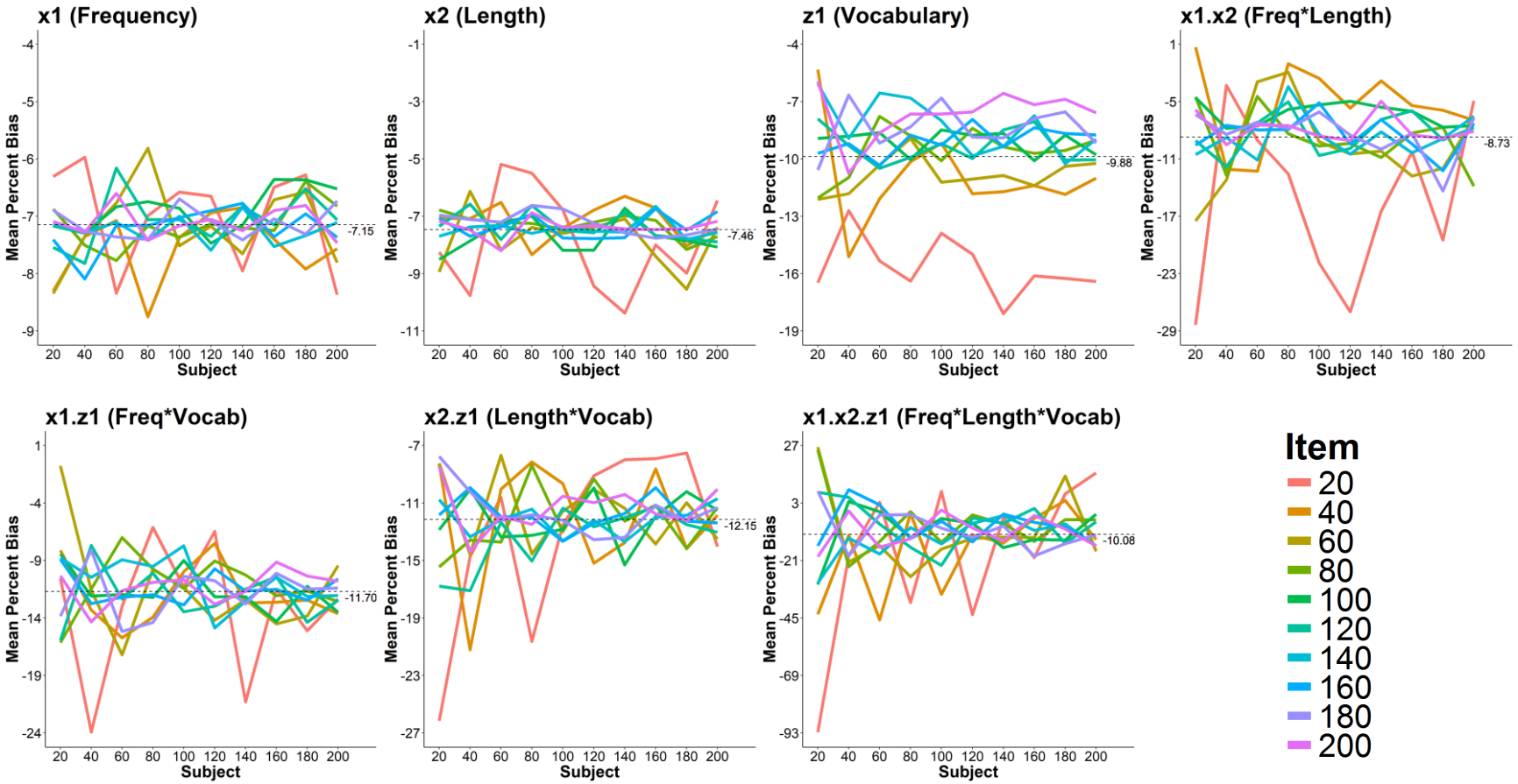

Figure 33. Average bias for each of the fixed-effect estimates in the raw model as a function of subject and item sample sizes. Simulation based on the model fit to the raw ELP data.

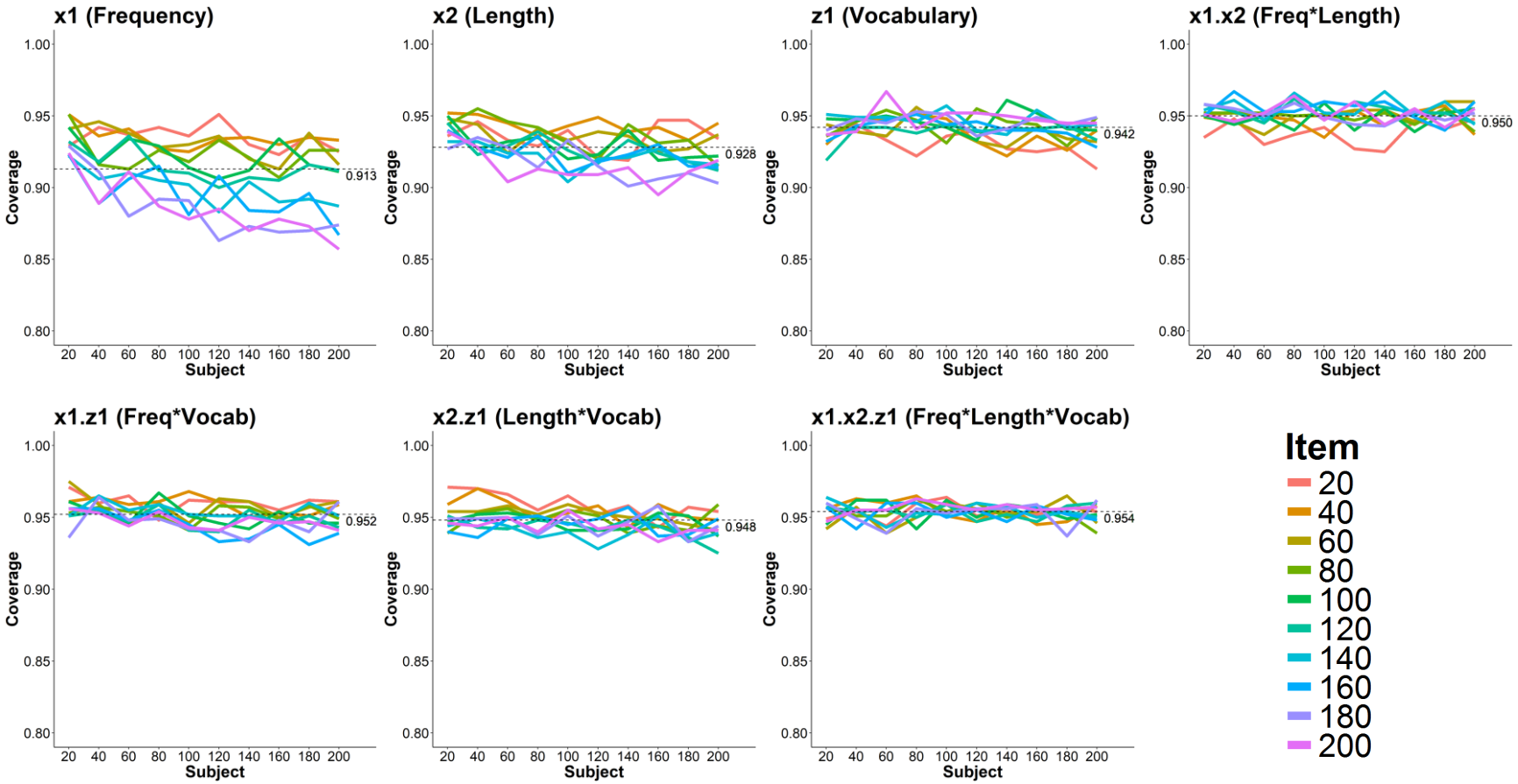

Figure 34. Coverage for each of the fixed effects in the raw model as a function of subject and item sample sizes. Simulation based on the model fit to the raw ELP data.

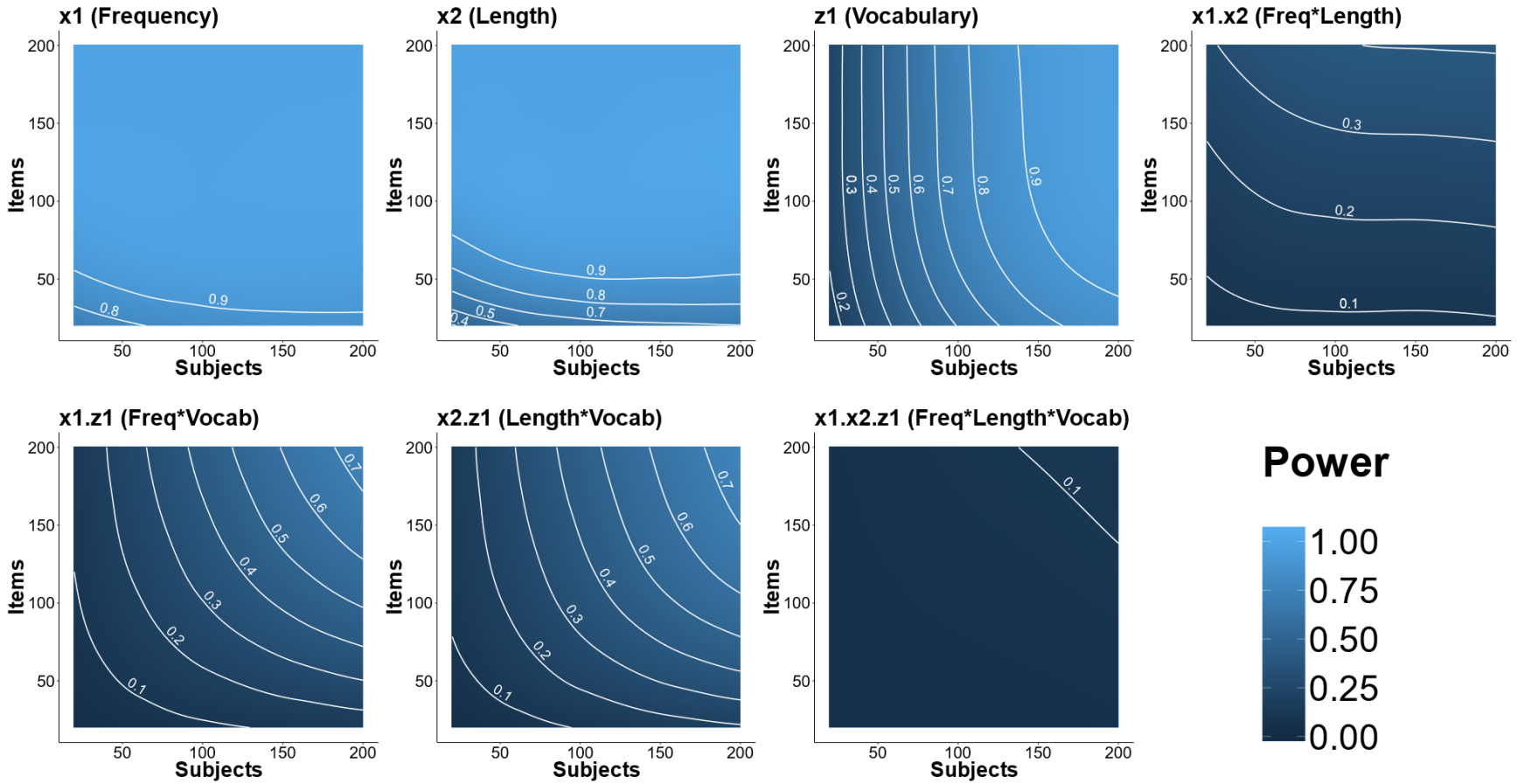

Figure 35. Power contour plots for the raw model as a function of subject and item sample size. Simulation based on the model fit to the raw ELP data.

Comparison measures: Transformed vs. raw model. Figure 36 shows the proportion of datasets where both the raw and transformed model converged as a function of subject and item sample sizes. As in Studies 1 and 2, item sample sizes appear to have a greater effect on this convergence consistency, such that when the sample had fewer items, models that converged when fit to raw data tended not to converge when fit to transformed data.

Subsequent comparisons between the raw and the transformed models were then made on datasets where both the raw and transformed model converged.

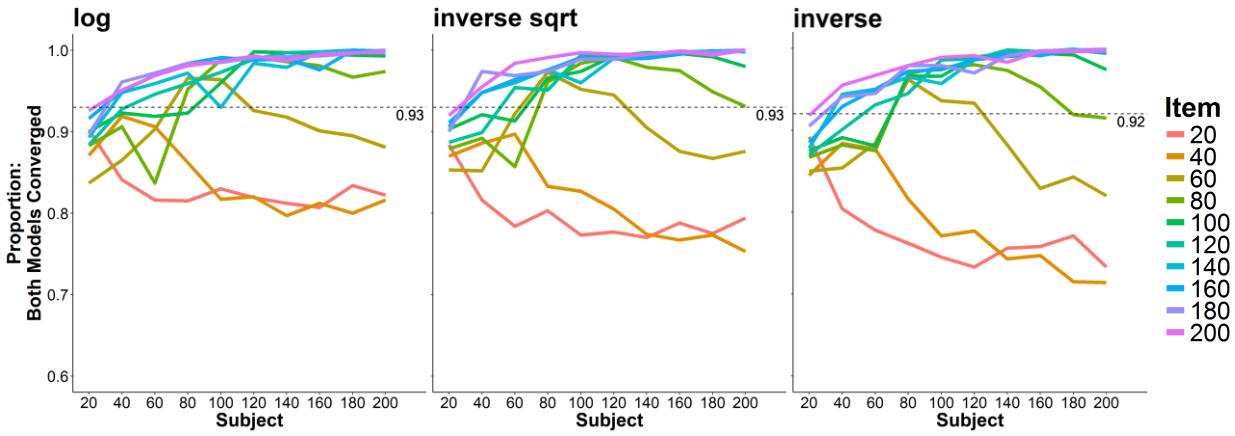

Figure 36. Proportion of datasets where the transformed models converged given that the raw model also converged, expressed as a function of subject and item sample sizes. Simulation based on the model fit to the raw ELP data.

Log model. The log model converged in 93,492 of the 100,000 datasets where the raw model converged. Increasing subject and item sample sizes also increased power in the log model for all effects except the x_1x_2 and three-way interactions as shown in Figure 37.

The contour plots in Figure 38 show the estimated differences in power between the raw and the log model for each of the effects as a function of subject and item sample sizes. For the main effects, the difference in power between the two models remained below 10% across sample sizes. However, for the x_1x_2, x_1z_1, and x_2z_1 two-way interactions, the *raw* model was estimated to have as much as 34%, 38%, and 18% more power than the log model as sample sizes increased respectively. The estimated difference for the three-way interaction remained below 6% across sample sizes.

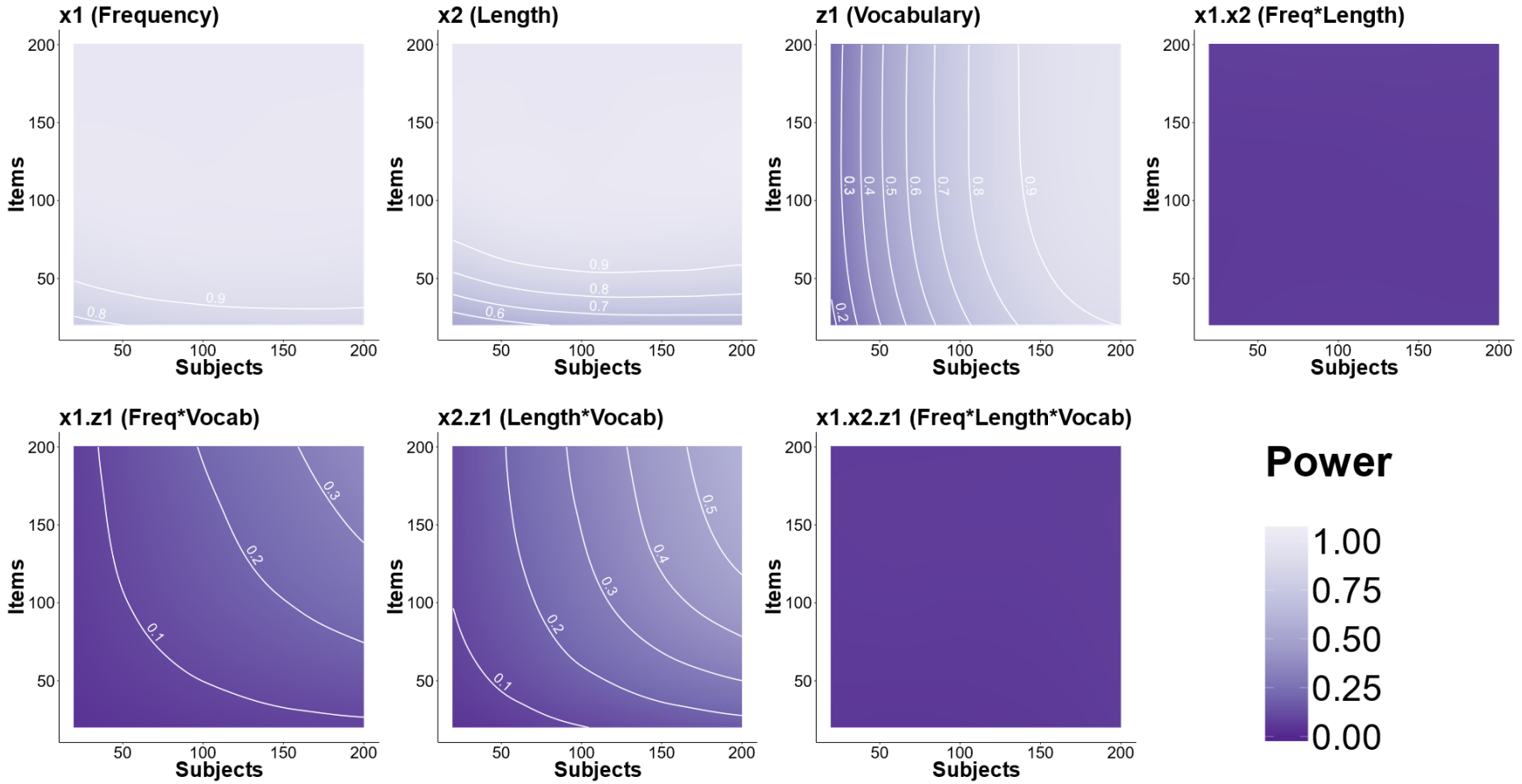

Figure 37. Power contour plots for the log model as a function of subject and item sample size. Simulation based on the model fit to the raw ELP data.

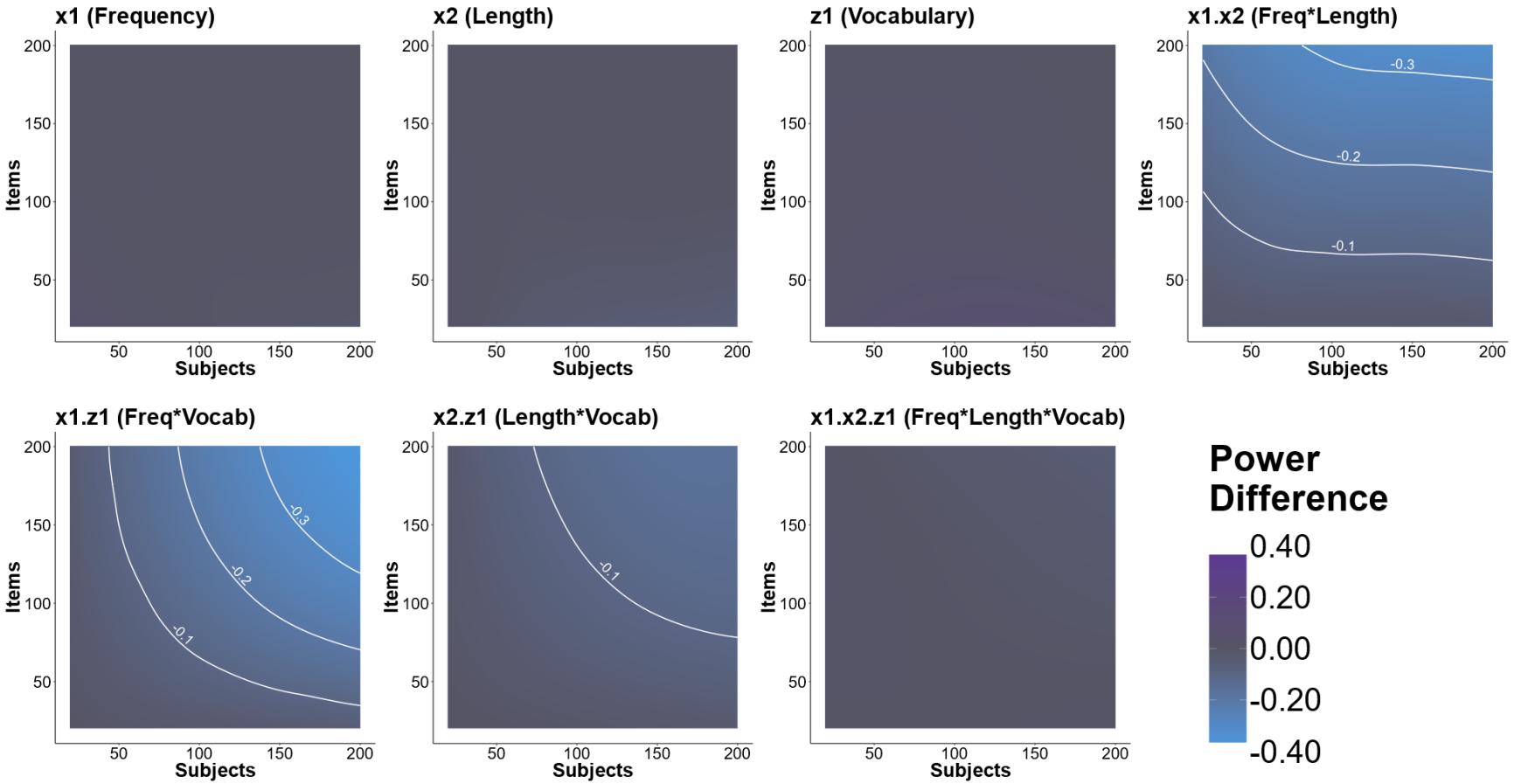

Figure 38. Power difference contour plots (log – raw) as a function of subject and item sample size. Blue indicates that the raw model has more power than the log model. Purple indicates that the log model has more power than the raw model. Simulation based on the model fit to the raw ELP data.

Inverse square-root model. The inverse square-root model converged in 92,976 of the 100,000 datasets where the raw model converged. Increasing subject and item sample sizes increased power for all effects except the x_1x_2 and x_1z_1 interactions as shown in Figure 39.

Figure 40 shows the estimated differences in power between the raw and the inverse square-root model for each of the effects as a function of subject and item sample sizes. For the main effects, the estimated difference in power between the two models remained below 12% across sample sizes. However, for the x_1x_2, x_1z_1, and x_2z_1 interactions, the *raw* model had as much as 37%, 62%, and 34% more power than the inverse square-root model as sample sizes increased respectively. This indicates that the inverse square-root model was worse than the log and raw models in detecting the two-way interactions despite being a stronger power transform. Lastly, the estimated difference for the three-way interaction remained below 4% across sample sizes. Recall that that the inverse square-root transform optimally normalized the raw ELP data.

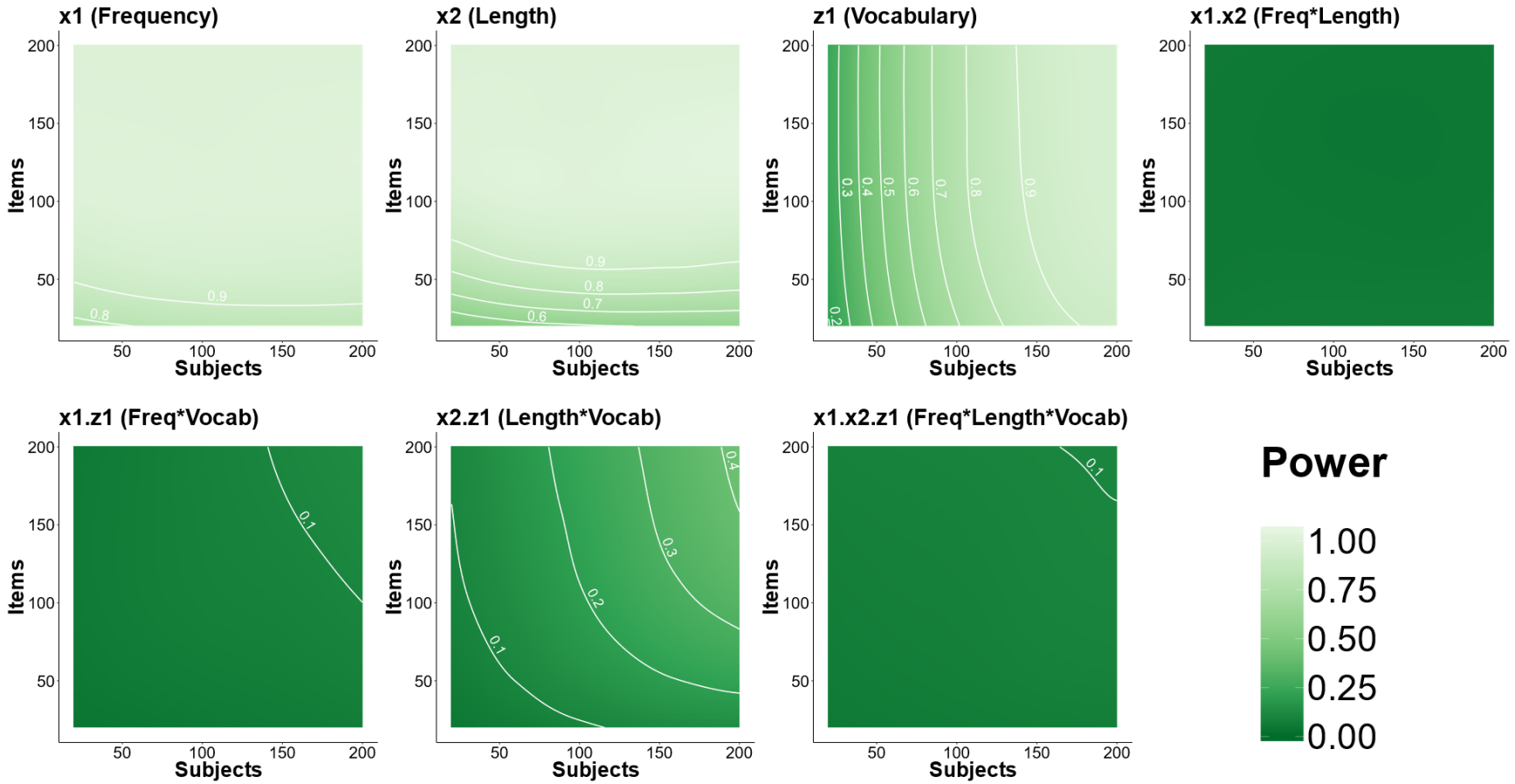

Figure 39. Power contour plots for the inverse square-root model as a function of subject and item sample size. Simulation based on the model fit to the raw ELP data.

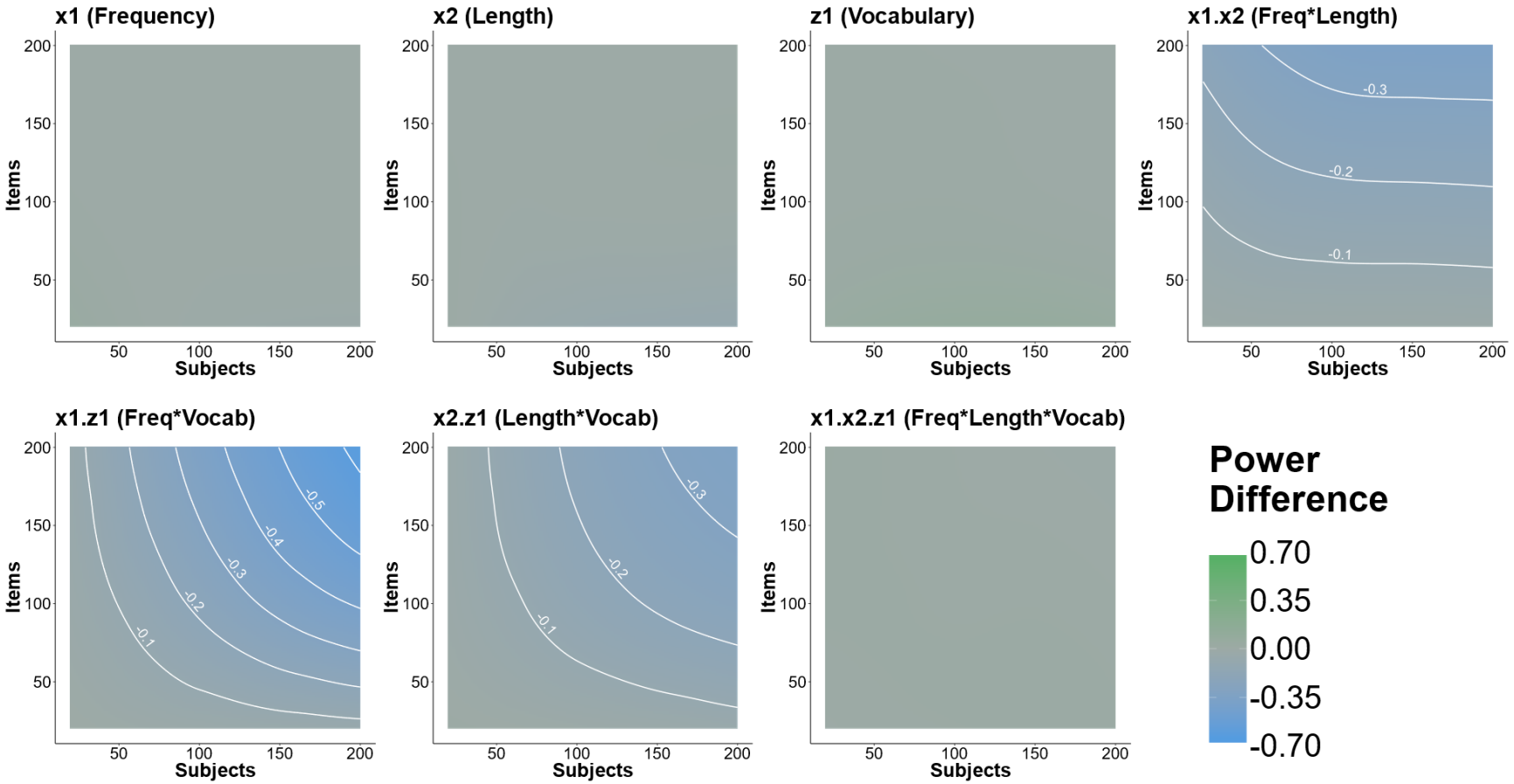

Figure 40. Power difference contour plots (inverse square-root – raw) as a function of subject and item sample size. Blue indicates that the raw model has more power than the inverse square-root model. Green indicates that the inverse square-root model has more power than the raw model. Simulation based on the model fit to the raw ELP data.

Inverse model. The inverse model converged in 91,660 of the 100,000 datasets where the raw model converged. Figure 41 shows that power increased as sample sizes increased for all effects, except the x_1x_2 and x_2z_1 interactions.

Figure 42 shows the estimated differences in power between the raw and the inverse model for each of the effects as a function of subject and item sample sizes. For the main effects, the estimated difference in power between the two models remained below 15% across sample sizes. However, for the x_1x_2, x_1z_1, and x_2z_1 interactions, the *raw* model was estimated to have as much as 37%, 67%, and 48% more power than the inverse square-root model as sample sizes increased respectively. The estimated difference for the three-way interaction remained below 7% across sample sizes.

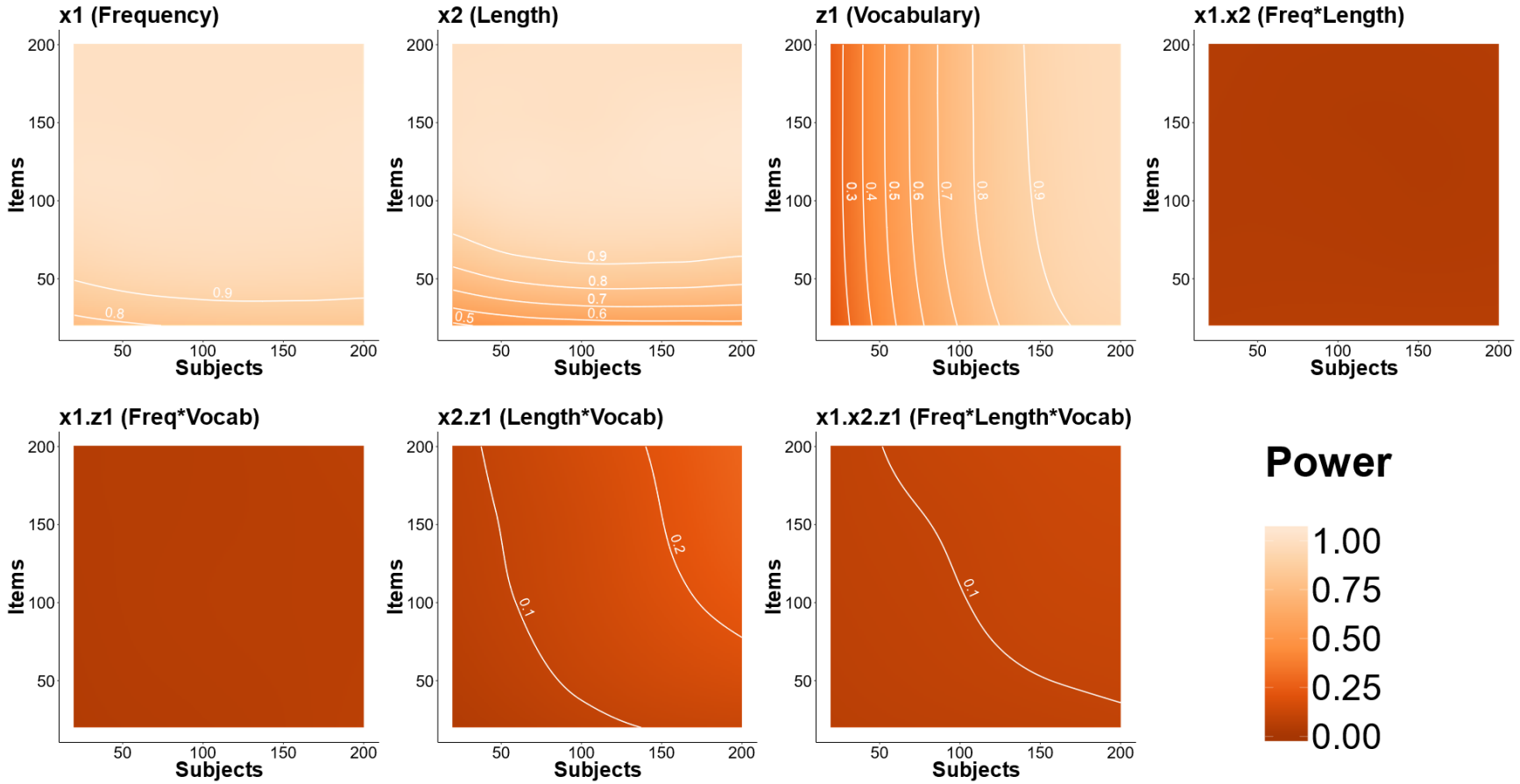

Figure 41. Power contour plots for the inverse model as a function of subject and item sample size. Simulation based on the model fit to the raw ELP data.

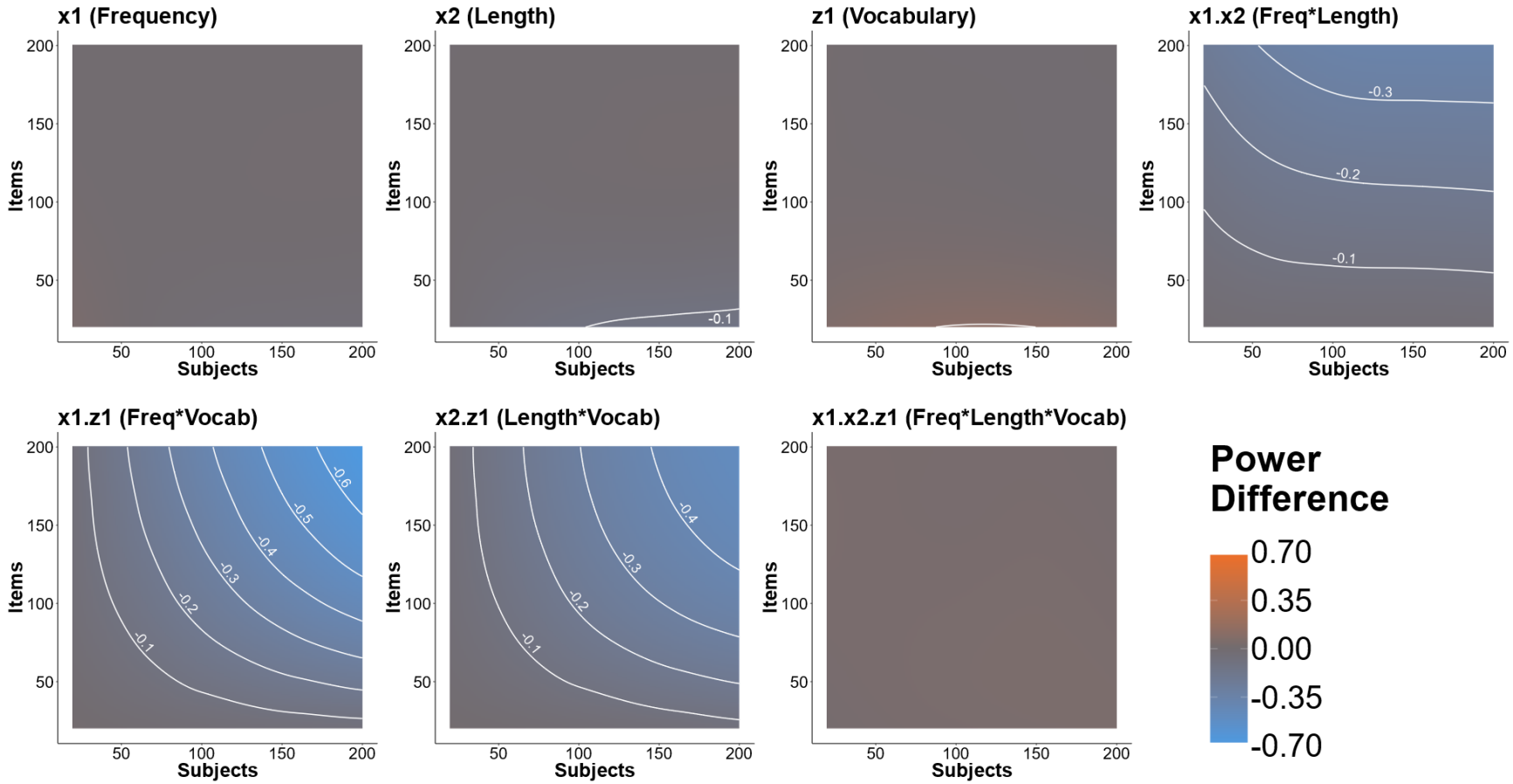

Figure 42. Power difference contour plots (inverse – raw) as a function of subject and item sample size. Blue indicates that the raw model has more power than the inverse model. Orange indicates that the inverse model has more power than the raw model. Simulation based on the model fit to the raw ELP data.

Summary. The results of this simulation are largely consistent with the results of Studies 1.2 and 2.2. The raw model underestimated all the effects and produced conservative 95%-confidence intervals for two of the main effects. Despite these consequences, it was only in the raw model – the statistically inappropriate model – that increases in sample sizes increased power for *all* of the effects in the model. All transformed models had two interaction effects for which increases in sample sizes did not increase power.

There were no systematic differences between the raw and the transformed models in detecting the main effects and the three-way interaction. But remarkably, as the power transform became stronger, the transformed model became substantially *worse* in detecting all two-way interactions as sample sizes increased. In fact, the raw model consistently outperformed the transformed models in detecting the interactions despite underestimating the effects.

As in the prior studies, I performed another simulation to address the possibility that these results were dependent on the generating process being based on the raw model. Because the Box-Cox procedure indicated that the optimal transform of the ELP data was the inverse square-root transform, the following simulation was based on the inverse square-root model fit to the ELP.

Study 3.3: Simulations based on inverse square-root model estimates

Generating process. The parameters used in the generating process of the simulation include all fixed- and random-effect estimates obtained from the inverse square-root model (see Table 9). With the Box-Cox procedure revealing that the inverse square-root transform optimally normalizes the RT data, a normal distribution with mean = 0 and variance = $4.4688^2 = 19.9706$ was used to generate trial-level residuals.

Dataset generation and analysis. This was identical to Study 3.2, except that the simulated inverse square-root RTs were back-transformed only into raw RTs to compare the performance of the models fit to these two RT scales.

Comparison measures: Inverse square-root vs. raw model. The same performance measures were obtained as in Study 1.2.

Results: Study 3.3

Figure 43 shows the proportion of datasets for which both the raw and inverse square-root models converged as a function of subject and item sample sizes. As in the prior simulations, models that converged when fit to the raw data did not necessarily converge when fit to the inverse-square-root-transformed data. Smaller item sample sizes seemed to exacerbate this issue. Subsequent results were evaluated on the 92,661 datasets where inverse square-root model also converged, out of 100,000 datasets where the raw model converged.

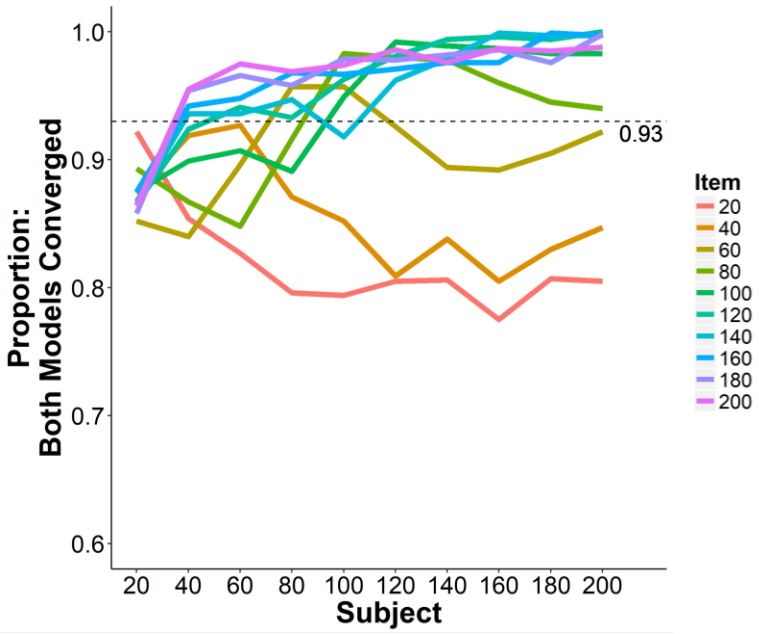

Figure 43. Proportion of datasets where the inverse square-root models converged when the raw model also converged, expressed as a function of subject and item sample sizes. Simulations based on the model fit to inverse-square-root-transformed ELP data.

The power contour plots for the raw model and inverse square-root model in Figures 44 and 45 show the general increase in power as subject and item sample sizes increase. As in Study 3.2, the estimated power for the x_1x_2 interaction did not reach 10% at the largest sample size condition for the inverse square-root model. The estimated power for the three-way interaction also did not reach 10% at the largest sample size for both the raw and inverse square-root model.

Differences between the raw and inverse square-root models' power estimates are shown in the power difference contour plots in Figure 46. While differences in power estimates for the main effects and the three-way interaction across all sample size conditions remained 7% or lower, the raw model's power for the x_1x_2, x_1z_1, and x_2z_1 interactions became as much as 16%, 48%, and 27% higher than the inverse square-root model's respectively as sample sizes increased. Again, the raw model outperformed the inverse-square root model despite the fact that the simulation's generating process was based on the inverse-square-root-transformed ELP data.

Overall, the results are similar to those observed in Study 3.2, thereby replicating Study 2 (FPP) and providing further evidence that the choice of underlying generating process does not change the pattern of results obtained from the simulations.

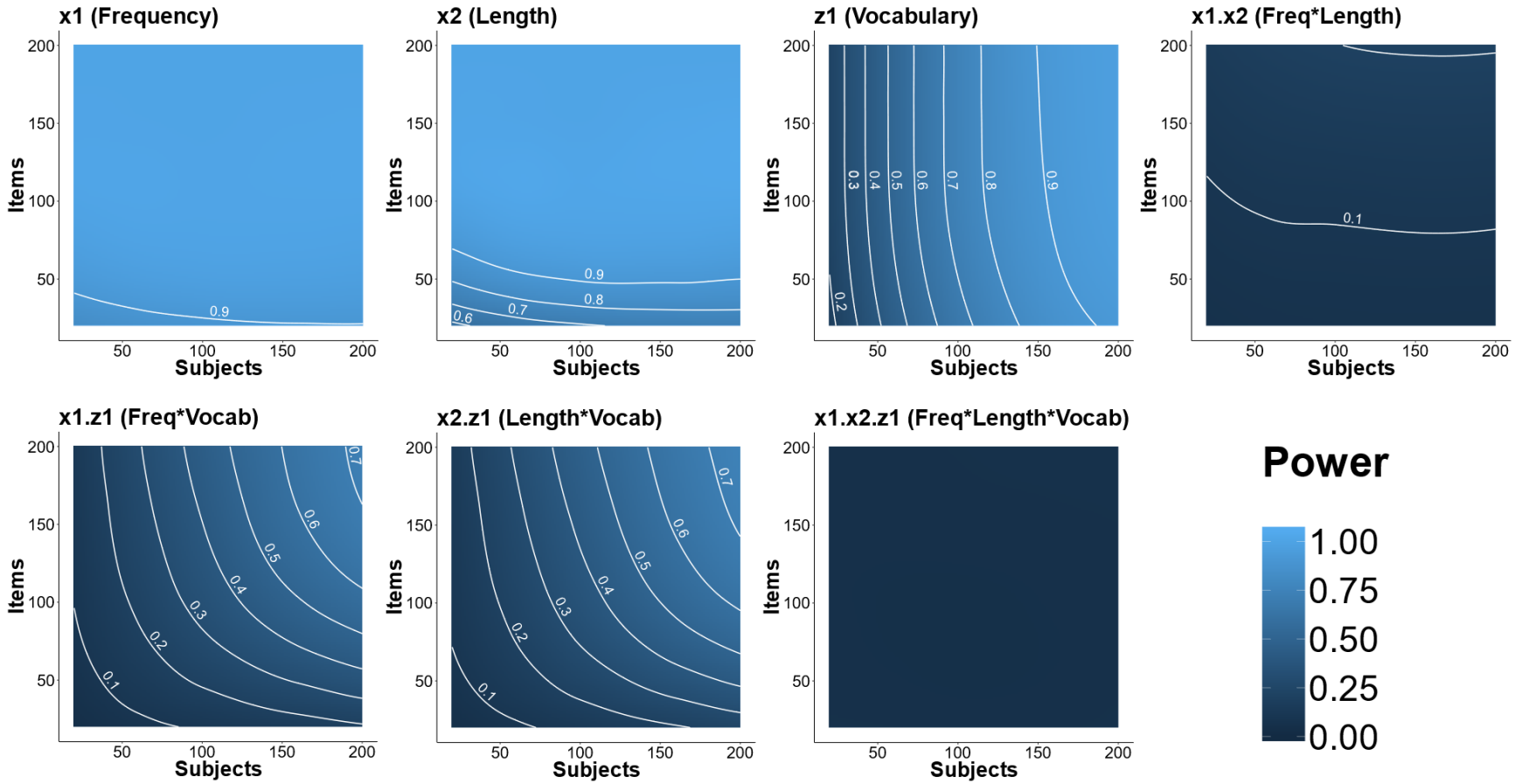

Figure 44. Power contour plots for the raw model as a function of subject and item sample size. Simulation based on the model fit to the inverse-square-root-transformed ELP data.

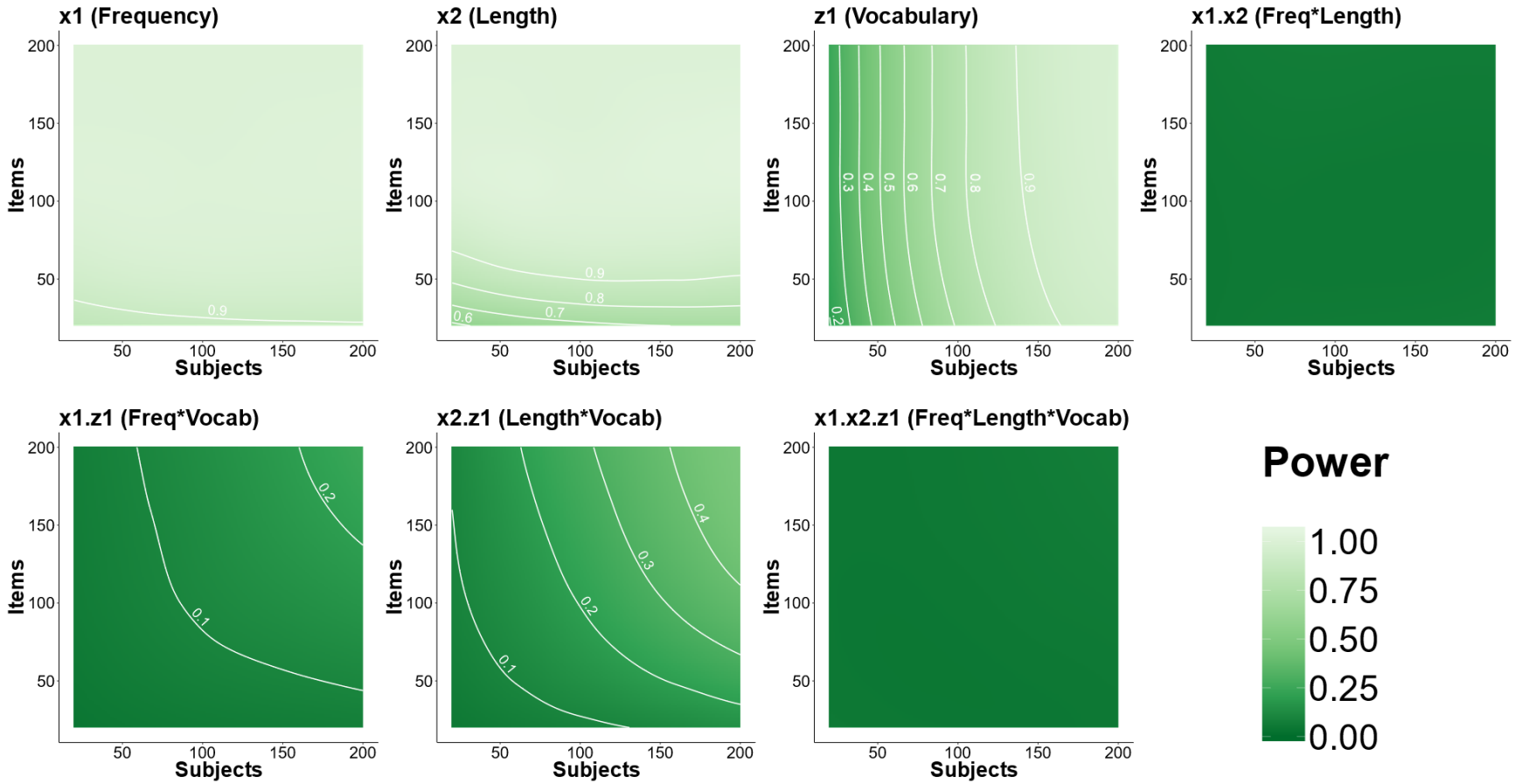

Figure 45. Power contour plots for the inverse square-root model as a function of subject and item sample size. Simulation based on the model fit to the inverse-square-root-transformed ELP data.

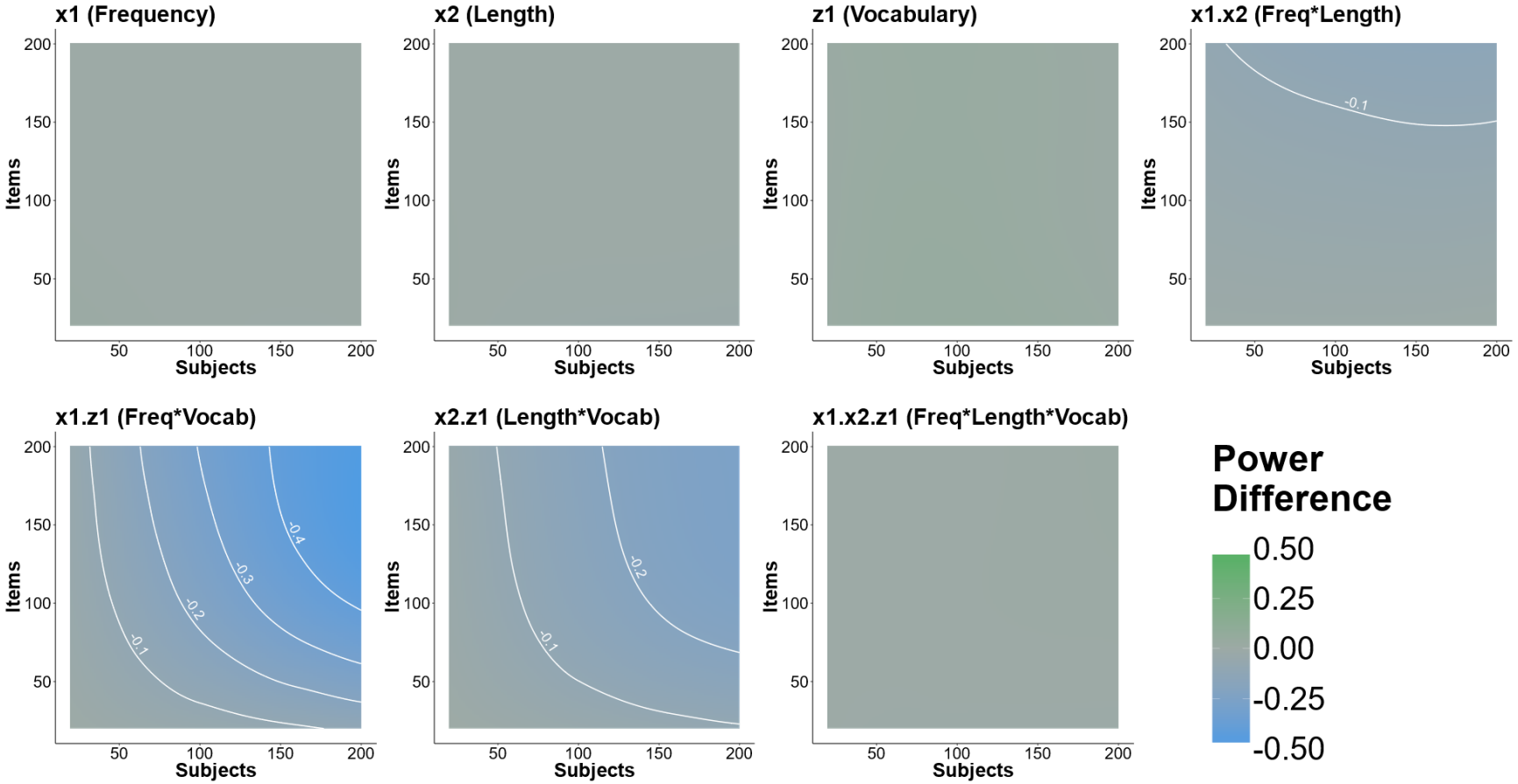

Figure 46. Power difference contour plots (inverse square-root – raw) as a function of subject and item sample size. Blue indicates that the raw model has more power than the inverse square-root model. Green indicates that the inverse square-root model has more power than the raw model. Simulation based on the model fit to the inverse-square-root-transformed ELP data.

General Discussion

In a series of three studies, I examined the influence of power transforms in LMMs fit to RT data. Each study consisted of three parts. In Part 1, I analyzed megastudy data by applying several power transforms to the RT data and fitting LMMs to each RT scale. I identified the transform that optimally normalized the residuals using the Box-Cox procedure and compared the fixed- and random-effect results between the raw and transformed LMMs. In Part 2, I simulated RTs using the LMM fit to the raw megastudy data as the generating process. I assessed how much bias and change in coverage are expected to be incurred from violating the normality assumption in fitting LMMs to raw RTs. I also compared the power and Type I error rates for main and interaction effects between the raw and transformed LMMs. In Part 3, I simulated RTs using the LMMs fit to the optimally transformed megastudy data as the generating process to ensure that the results obtained in Part 2 did not depend on using the raw LMM to generate the simulations. As in Part 2, I compared the power and Type I error rates for main and interaction effects between the raw and transformed LMMs. The key results for each of these parts across the three megastudies are summarized in Table 10.

Results	Study 1 Semantic Priming Project (SPP) Hutchison et al. (2013)	Study 2 Form Priming Project (FPP) Adelman et al. (2014)	Study 3 English Lexicon Project (ELP) Balota et al. (2007)
Part 1: Analysis of megastudy data			
Optimal (Box-Cox) power transform	• λ = -0.95, approximating the inverse transform	• λ = -0.71, which is closer to the inverse square-root than the inverse transform	• λ = -0.46, approximating the inverse square-root transform
Influence on main effects	• *t*-statistics increased for x_1, x_2 and z_1	• *t*-statistics slightly decreased for x_1, increased for x_2 and z_1	• *t*-statistics increased for x_1 and x_2; negligible change for z_1
Influence on same-level interactions (x_1x_2)	• *t*-statistic for x_1x_2 decreased	• *t*-statistic for x_1x_2 decreased	• *t*-statistic for x_1x_2 decreased
Influence on cross-level interactions (x_1z_1, x_2z_1, $x_1x_2z_1$)	• *t*-statistic for x_1z_1 decreased, x_2z_1 increased past critical value for significance	• *t*-statistics for x_1z_1 and x_2z_1 decreased	*t*-statistics for x_1z_1 and x_2z_1 decreased; $x_1x_2z_1$ reversed direction
Influence on random effects	• Reversal of correlation between subjects' RT intercepts and absolute magnitude of their word frequency slopes from *r* = 0.48 to *r* = -0.36	• Reversal of correlation between subjects' RT intercepts and absolute magnitude of their word frequency slopes from *r* = 0.49 to *r* = -0.11	• Reduction in correlation between subjects' RT intercepts and absolute magnitude of their word frequency slopes from *r* = 0.73 to *r* = 0.33
Part 2: Simulations based on raw model as generating process			
Bias in raw LMM	• All nonzero fixed effects other than x_2 were underestimated	• All nonzero fixed effects other than x_2 were underestimated	• All fixed effects were underestimated
Coverage in raw LMM	• x_1 had conservative coverage	• x_1 had conservative coverage	• x_1 and x_2 had conservative coverage
Convergence	• Transformed LMMs converged in 88% to 90% of datasets where raw LMM converged	• Transformed LMMs converged in 90% to 91% of datasets where raw LMM converged	• Transformed LMMs converged in 92% to 93% of datasets where raw LMM converged.
Relative power/Type I error for transformed vs. raw LMMs			

Main effects	• Relative power increased for all main effects in small sample sizes	• Relative power increased for x_1 and x_2 in small sample sizes; unaffected for z_1	• Relative power unaffected for all main effects
Same-level interactions	• Relative power increased for x_1x_2 as sample sizes increased	• Relative power decreased for x_1x_2 as sample sizes increased	• Relative power decreased for x_1x_2 as sample sizes increased
Cross-level interactions	• Relative power and Type I error unaffected for cross-level interactions	• Relative power and Type I error unaffected for cross-level interactions	• Relative power decreased substantially for x_1z_1 and x_2z_1 as sample sizes increased; unaffected for $x_1x_2z_1$
Part 3: Simulations based on transformed model as generating process			
Relative power/Type I error for transformed vs. raw LMMs			
Main effects	• Relative power increased for x_1 and x_2 in small sample sizes; unaffected for z_1	• Relative power increased slightly for x_2 in small sample sizes; unaffected for x_1 and z_1	• Relative power unaffected for main effects
Same-level interactions	• Relative power decreased for x_1x_2 as sample sizes increased	• Relative power decreased substantially for x_1x_2 as sample sizes increased	• Relative power decreased slightly for x_1x_2 as sample sizes increased
Cross-level interactions	• Relative power decreased substantially for x_1z_1 as sample sizes increased and increased slightly for x_2z_1; Type I error unaffected for $x_1x_2z_1$	• Relative power decreased for x_1z_1 as sample sizes increased; Type I error unaffected for x_2z_1 and $x_1x_2z_1$	• Relative power decreased substantially for x_1z_1 and x_2z_1 as sample sizes increased; unaffected for $x_1x_2z_1$

Table 10. Summary of results from all three studies in the current project.

Part 1: Analyzing the megastudies. Consistent with previous demonstrations in smaller datasets (Balota, Aschebrenner, & Yap, 2013; Lo & Andrews, 2015), analyses of the megastudies revealed that that power transforms alter interaction patterns observed in the raw scale. Power transforms either had no effect or magnified the *t*-statistics observed for the main effects compared to the raw model. However, for the interaction effects, stronger power transforms were associated with systematic changes in *t*-statistics compared to the raw model, so much so that the *t*-statistics for some of the interaction effects became significant (x_2z_1 for Study 1: SPP), became nonsignificant (x_1z_1 for Study 3: ELP), or even changed signs ($x_1x_2z_1$ in Study 3: ELP) for the inverse-transformed data.

Beyond the fixed effects, the power transforms also altered random-effect correlation patterns present in the raw scale. In some cases, the transformed models estimated weaker relationships between the random effects than the raw model; in Study 3 (ELP), bigger word frequency effects were strongly associated with slow subjects in the raw model, whereas this association was only modest in the inverse-square-root model. However, in other cases, the transformed models estimated *completely different* relationships between the random effects from those estimated in the raw model. In Study 1 (SPP), bigger word frequency effects were associated with slow subjects in the raw model (as in Studies 2 [FPP] and 3 [ELP]), but *smaller* frequency effects were associated with slow subjects in the optimally transformed models.

That power transforms also alter random-effect patterns is notable because psychologists' discussions about the statistical implications of power transforms thus far have only focused on fixed effects. This is partly because the discussions have only involved ANOVAs until recently (e.g., Balota, et al., 2013; Levine & Dunlap, 1982; Lo & Andrews, 2015; Loftus, 1978; Ratcliff, 1993; Wagenmakers et al., 2012) but also because psychologists have been primarily interested in making inferences about their manipulations and predictors of interest. However, random effects lend insight into potential systematic interindividual variability in the effects we observe, thereby providing information we can use to enrich the theories and models we develop about those effects. There is little reason to believe that manipulations or predictors induce effects of some true exact magnitudes and that deviations from these

values are necessarily aberrations (Speelman & McGann, 2013). Even if the effect of a manipulation might be in the same direction for all subjects, there could very well be individual differences in the strength of that effect due to factors that impact subjects' cognitive architectures. Therefore, it is worthwhile to examine how power transforms might affect the estimation of random-effect patterns. Since random slopes and intercept-slope correlations are interactions yet to be accounted for, changes in these random effects could indicate that the potential interactions they absorbed would also be altered by power transforms.

Part 2: Simulations based on the raw model as generating process. Overall, the simulations based on the raw models fit to the megastudies reinforced the results from the megastudy data analyses and revealed the following:

Bias. In all three studies, the raw LMM tended to underestimate all the effects specified in the model except for x_2 in Study 2.2 (FPP), though the magnitude of the bias varied widely as a function of subject and item sample size across the studies.

Coverage. The raw model also had consistently conservative coverage for main effects. In Studies 1: SPP and 2: FPP, the average coverage for the word frequency effect (x_1) was conservative at 93.1% and 91.9% respectively; in Study 3: ELP, the average coverage for all the main effects were conservative from 91.2% to 92.9%. As with the results on bias, coverage also varied widely as a function of subject and item sample size across the three studies.

***Convergence*.** On average, the transformed LMMs converged in 88 to 94% of the simulated datasets where the raw LMM also converged across the three studies. Of the datasets where the two models did not consistently converge, most consisted of few subjects and items, with item sample size exerting greater influence on consistent convergence. Note that the convergence rates obtained from the current simulations – and therefore the extent to which both the raw and transformed LMMs would converge in the same datasets – are likely only generalizable to LMMs of similar complexity and represent a lower bound for simpler models. Lower convergence rates are expected for more complex models, particularly for maximal models (Barr et al., 2013).

Besides random effects, model nonconvergence should also be considered in discussions of the implications of power transforms in LMMs fit to RT data. Wagenmakers et al. (2012) recommend fitting statistical models to multiple scales to determine whether results are consistent across scales. But the simulation results revealed that models that converge in one scale would not necessarily converge in another scale. When this situation occurs, it not only invites ad hoc decisions to analyze the data in the scale in which the LMM converged, but it also impedes opportunities for further model scrutiny. Should there be differences between the raw and transformed models, the differences cannot be identified and evaluated because one of the models is untrustworthy due to nonconvergence. Scale selection should occur prior to data analysis and model selection and random effect specification should be performed exclusively on the selected scale (see further below).

***Power and Type I error*.** Despite the consequences mentioned earlier, the power of the raw model to detect main effects is still comparable to those of the transformed as subject and item sample sizes increased. The transformed models tended to be more sensitive in detecting main effects in samples with fewer subjects and fewer items, consistent with simulations examining the influence of power transforms in samples of this size in the context of ANOVAs (Levine & Dunlap, 1982; Ratcliff, 1993). But the difference in power between the raw and transformed models was eliminated as sample sizes increased.

The raw model also tended to be *more sensitive* in detecting existing interaction effects than the optimally transformed model, especially as sample sizes increased. The only situation in which the optimally transformed model had more power than the raw model was in detecting the interaction of word frequency and vocabulary (x_1z_1) in Study 1.2. For null interaction effects, the raw and transformed models had comparable Type I error rates, which were around the nominal 5% rate (i.e., Type I error rates never exceeded 8%) regardless of subject and item sample sizes. The effects of power transforms on statistical power and Type I error were similar for same-level (x_1x_2) and cross-level interactions (x_1z_1, x_2z_1, $x_1x_2z_1$).

Part 3: Simulations based on the optimally transformed model as the generating process. Similar results on convergence and power/Type I error were obtained when the simulations were

generated using the optimally transformed models fit to the megastudies. This was unexpected: if the generating process was based on the optimally transformed model, it is reasonable to expect the optimally transformed model to perform better than the raw model. This was not the case: in fact, the raw model still had greater power in detecting most of the existing interactions than the optimally transformed model in all three studies when the generating process was based on the optimally transformed model. The only exception was the interaction between semantic priming and vocabulary ($x_2 z_1$) in Study 1.3: SPP, which the inverse model was slightly more powerful in detecting. This indicates that the results of the simulations did not depend on the generating process: even when the functional form of the relationships between the predictors and RT is *not* linear, the model that assumes linear relationships between the predictors and RT (i.e., the raw model) was still more powerful in detecting interactions than the very model that assumes the exact nonlinear functional form (i.e., the optimally transformed model).

Summary. Collectively, the results indicate that that the influence of power transforms on LMMs generalize across subject and item sample sizes and across different kinds of studies. Despite making models meet the normality assumption more closely, stronger power transforms led to lower rates of detecting interaction effects present in both the raw and transformed scales. This is analogous to the systematic shrinking of the interaction effects' *t*-statistics observed from analyzing the megastudies (Part 1). Moreover, despite incurring statistical consequences from violating the normality assumption, the raw model detected main effects at a similar rate as the transformed models and tended to be *more sensitive* in detecting existing interaction effects as subject and item sample sizes increased. This was regardless of whether the raw model (Part 2) or the optimally transformed model (Part 3) was the data-generating model in the simulations.

Transforming RTs in Psycholinguistics

Strictly speaking, statistical models need to meet their respective assumptions in order for inferences made on them to be valid. When statisticians anticipate violating these assumptions based on the structure and properties of their data, their approach to preempting the violation of these assumptions

fall into two general categories: they change the model to fit the data or they change the data to fit the model. Of the latter, a typical procedure has been to apply power transforms to skewed data in order to avoid violating the normality assumption and to tame outlier observations (e.g., Snijders & Bosker, 2012; Raudenbush & Bryk, 2002). When LMMs were introduced in psycholinguistics (e.g., Baayen, 2008; Baayen, Davidson, & Bates, 2008; Baayen & Millin, 2010), researchers openly adopted this procedure from statisticians to analyze chronometric data. In recommending to transform RTs, Baayen (2008) reasoned that:

> "[RTs are transformed] to eliminate or at least substantially reduce the skewing in their distribution. This reduction is necessary for most of the statistical techniques…to work appropriately. Without the…transformation, just a few outliers might dominate the outcome, partially or even obscuring the main trends characterizing the majority of data points." (p. 31)

Since then, power transforms have been regularly used when LMMs are fit to chronometric data. A manual survey of articles in the *Journal of Memory and Language* from January 2013 to December 2017 using "Baayen, Davidson, & Bates (2008)" or "Baayen (2008)" or "lme4" and "reaction time" or "response time" or "fixation duration" as search terms revealed that 31 (46%) out of the 68 articles where LMMs were fit to chronometric data involved the use of power transforms. In all 31 articles, researchers justified the use of power transforms simply to reduce the impact of outliers, reduce the positive skew, or to preempt violating the normality assumption, though interestingly, only 7 (23%) of the articles reported diagnostic measures (e.g., QQ plots, Box-Cox procedure, tests for normality) that indicated that normality was optimally approximated by the power transform that was applied.

However, the results of the current project demonstrate that applying power transforms to chronometric data tends to decrease the LMM's sensitivity to detect interaction effects. The pertinence of these results is underscored by 62 (91%) of the 68 articles mentioned above that examined interactions. Clearly, in cases where researchers applied power transforms and failed to find their interaction(s) of interest, we cannot determine whether these failures are due to the absence of the effect in the population, the study being underpowered in detecting the interaction effect, the decreased sensitivity of the transformed LMM, or a combination of these reasons. Nevertheless, that these failures are possibly due to

the decreased sensitivity of the transformed LMM warrants more careful consideration of using power transforms when fitting LMMs to chronometric data.

Transforming RTs in Cognitive Psychology

The widespread uncritical use of power transforms in LMMs is surprising given the substantial discussion on their implications for interactions within other areas of cognitive psychology. As discussed earlier, the idea that power transforms potentially influence conclusions about the presence of interactions in the raw scale is not at all new. Sternberg (1969) claimed that power transforms will generally destroy patterns of additivity in the raw scale; Loftus (1978) argued that only a privileged set of interactions – interactions that cross – are unaffected by power transforms, and all else are scale-dependent or removable interactions. This idea was revived in the context of LMMs in recent reanalyses by Balota, et al. (2013) and Lo & Andrews (2015), but as noted by Wagenmakers et al., (2012), many researchers are still unaware that the presence of interactions can depend on the scale of analysis. By analyzing data from several megastudies and performing multiple simulations, the current project extends and qualify this body of literature. Specifically, power transforms decrease the sensitivity of LMMs to detect existing interactions and do not change the Type I error rates of LMMs for interactions.

It can be argued that the power transforms did not decrease the sensitivity of LMMs to detect interactions *per se*, but rather the inability of transformed LMMs to detect interactions is a byproduct of the scale-dependence of all the tested interactions. Specifically, it is possible that in the population, the tested interactions do not manifest in some of the transformed scales that were used, which therefore prevented the LMMs fit to RT in those scales from detecting the interactions. After all, the systematic changes in *t*-statistics across the two-way interactions as stronger power transforms were used in Part 1 of all three studies suggest that if not for the size of the megastudies, some interactions in certain transformed models might not have reached significance in the first place (e.g., x_2z_1 in Study 1.1: SPP and x_1z_1 in Study 3.1: ELP). Thus, if these scale-dependent interactions do not manifest in certain

transformed scales in the population, the inability of transformed LMMs to detect these scale-dependent interactions in the simulations should be unsurprising, if not expected.

While this characterization is reasonable, the argument implies that LMMs fit to RT in scales where the scale-dependent interactions manifest in the population should have improved power in detecting the interactions, if not comparable to that observed in LMMs fit to raw RTs. However, Part 3 of all three studies critically demonstrated that even in scales where the scale-dependent interaction is present in the population (as specified in the simulations), LMMs fit to RTs in those very scales are still *less* sensitive in detecting the interactions than LMMs fit to raw RT. Thus, (optimally) transformed LMMs tend to have decreased sensitivity in detecting interactions regardless of the scale-dependence of the interactions in the population. I hypothesize that this decreased sensitivity is due to power transforms expanding the short end and compressing the long end of the RT distribution.

To Transform or Not to Transform? The Answer is in Theory, Not Normality

Given the theoretical and statistical consequences associated with applying power transforms to chronometric data, power transforms should not be used simply to meet the normality assumption. Part of their widespread use for this purpose might be due to the misleading impression that the power transform that optimally normalizes the residuals also reveals the scale in which the predictors and the RTs are related. But there is little reason to believe that requiring different power transforms to optimally normalize the RTs in each of the megastudies (i.e., inverse, inverse square-root) indicates that the relationships between the predictors and the RTs also take on different functional forms depending on the study. All three megastudies measured the same behavior with the same lexical decision task, which implies that the processes that underlie this behavior and the functional form of the relationships to be modeled are the same across the megastudies.

Fitting LMMs to both raw and transformed RTs may not also be effective in validating results obtained from both models (contra Wagenmakers et al., 2012). Inconsistent results between LMMs fit to different scales pose a conundrum about which of the LMMs should be interpreted. The current project

extends this concern to models that reveal *consistent* results. Consistency of LMMs fit to different scales is not a reassuring measure of the presence or scale-dependence of an interaction. Failing to detect an interaction in both raw and transformed LMMs could be due to the study being underpowered or the interaction being scale-dependent and not manifesting in the selected scales of analysis. On the other hand, detecting an interaction in both raw and transformed LMMs could prompt the incorrect conclusion that the detected interaction is scale-independent when the interaction just happens to manifest in both scales of analysis.

The decision to use to a power transform should be therefore motivated by a theoretical framework. When LMMs are fit to raw RTs, it is assumed that the underlying cognitive constructs measured by the predictors directly affect the time it takes to perform the mental operations that underlie the observed behavior. In contrast, transformed LMMs can be motivated by process models which posit that the outcome of interest is not time *per se* but a latent outcome tapped by time such as efficiency of processing. Process models that assume nonlinear mappings between the latent cognitive process, the latent outcome, and the observed RT, such as the diffusion model (Ratcliff, 1978; Wagenmakers et al., 2012), may therefore warrant transforming RTs. In both cases, the selected scale is associated with specific assumptions about the nature of the relationship between the predictors and the observed RT. Thus, researchers should motivate their scale of analysis prior to analyzing the data and acknowledge and address the specific statistical consequences associated with their selected scale as identified in the current project. Raw LMMs will tend to have biased fixed-effect estimates and conservative coverages for main effects, whereas transformed LMMs will have higher power for main effects in smaller sample sizes but will tend to be less sensitive than raw LMMs in detecting interactions regardless of the scale-dependence of the interactions.

If the raw scale is selected as the scale of analysis, the consequences incurred from violating the normality assumption may be addressed by estimating robust standard errors for the fixed effects (e.g., using the Huber-White sandwich estimator; Maas & Hox, 2004; Snijders & Bosker, 2012; Raudenbush & Bryk, 2002). A promising though more complicated alternative is to forgo the normality assumption and

acknowledge the positive skew of RT data in the population by fitting *generalized linear mixed-effects models* (GLMMs) (Lo & Andrews, 2015). However, because maximum likelihood estimates cannot be analytically determined in GLMMs, GLMMs are more difficult to fit than standard LMMs and fail to converge at higher rates for models that would be otherwise easy to fit with standard LMMs. Consequently, fitting maximal GLMMs are much less plausible, and GLMMs may have to be drastically modified to a point where potentially invalid assumptions about empirical patterns would have to be made for the model to converge (e.g., subjects' intercepts are *unrelated* to the magnitude of their word frequency effects; see Supplemental Material of Lo & Andrews, 2015), and this limitation is not necessarily offset by increasing sample sizes. In fact, I started the current project with the intention to demonstrate the advantage of GLMMs, but I was unable to pursue this plan due to this very issue. Thus, while GLMMs are promising in addressing the positive skew in chronometric data, their advantage is currently limited by the difficulty with which they are fit.

If a transformed scale (e.g., log, inverse) is selected as the scale of analysis, increasing sample sizes *might* address the decreased sensitivity in detecting interaction effects. Increasing sample sizes would be ineffective if the interactions of interest are scale-dependent and they do not manifest in the population in the scale selected for analysis. However, because it is unknowable whether the failure to detect a scale-dependent interaction is due to the interaction not manifesting in the selected scale of analysis or to the power transform decreasing the sensitivity of the LMM, increasing the sample size would provide the most information about the status of the tested interactions in the population.

Conclusion

Applying power transforms to chronometric data has conceptual and statistical implications that outweigh the recommendation to use them in LMMs to meet the normality assumption. They not only raise questions about the appropriate scale in which chronometric data should be analyzed, but they also affect LMMs in ways that are substantially more consequential than violating the normality assumption. Thus, using power transforms should not be driven by the need to meet modeling assumptions – which is

unfortunately how power transforms currently tend to be used in the literature – but by hypotheses regarding the functional form of the relationship between factors of interest and time. Under a framework where manipulations and predictors are thought to directly affect the duration of mental processes, raw RT is clearly the outcome of interest (Lo & Andrews, 2015). In another framework, time may be treated as an indirect measure of processing efficiency, and it is this latent outcome which unobserved mental constructs affect, most likely nonlinearly (Wagenmakers et al., 2013). In this case, process models may identify a transformation linking the factors that tap into these mental constructs and observed RT, thereby making transformed RT the outcome of interest.

Consequently, the scale of analysis should be theoretically motivated prior to analyzing the data and with consideration of the specific statistical consequences associated with fitting LMMs to that scale beyond those related to violating the normality assumption. The simulations showed that LMMs fit to transformed data will tend to be more sensitive in detecting main effects in small samples but less sensitive in detecting interaction effects as sample sizes increase than LMMs fit to raw data. Thus, a researcher who is interested in detecting an interaction and has no strong commitments about the scale in which factors of interest affect cognitive processing would gain power for the interaction by fitting an LMM to raw RT. On the other hand, a researcher who plans to fit an LMM to transformed RT would gain power in detecting main effects but would need to compensate for the decreased power in detecting interaction effects by increasing sample size. This tradeoff is greater when a stronger power transform is applied and a more complex model is fit to the data.

On interpreting results, Cohen (1994, p. 1001) said that "psychologists have to start respecting the units they work with, or develop measurement units they can respect enough so that researchers in a given field or subfield can agree to use them." There are very few measures in psychology whose units are respected enough so as to be agreed upon as important such as response time. The uncritical, widespread use of power transforms on chronometric data to preempt the violation of the normality assumption is a breach of this respect, and careful consideration of the conceptual and statistical implications of using power transforms should redeem it.

REFERENCES

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*(2), 12-28.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412.

Balota, D. A., & Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. *Journal of Experimental Psychology: General, 128*(1), 32-55.

Balota, D. A., & Yap, M. J. (2011). Moving beyond the mean in studies of mental chronometry the power of response time distributional analyses. *Current Directions in Psychological Science, 20*(3), 160-166.

Balota, D. A., Aschenbrenner, A. J., & Yap, M. J. (2013). Additive effects of word frequency and stimulus quality: The influence of trial history and data transformations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5), 1563-1571.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445-459.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-278.

Bates, D., Maechler, M., Bolker, B.M., & Walker, S. (20142015). lme4: Linear mixed-effects models using Eigen and S4. ArXiv e-print; submitted to *Journal of Statistical Software*. Retrieved from http://arxiv.org/abs/1406.5823Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48.

Berry, W. D., DeMeritt, J. H., and Esarey, J. (2010). Testing for interaction in binary logit and probit models: is a product term essential? *American Journal of Political Science, 54*, 248-266.

Borowsky, R., & Besner, D. (1993). Visual word recognition: a multistage activation model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(4), 813-840.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 211-252.

Brown, C., & Hagoort, P. (1993). The processing nature of the N400: Evidence from masked priming. *Journal of Cognitive Neuroscience*, *5*(1), 34-44.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977-990.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1), 1-20.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335-359.

Cohen, J. (1994). The earth is round (p<. 05). *American Psychologist*, *49*(12), 997-1003.

Donders, F.C. (1969). On the speed of mental processes. *Acta Psychologica*, *30*, 412-431.

Forster, K. I., Davis, C., Schoknecht, C., & Carter, R. (1987). Masked priming with graphemically related forms: Repetition or partial activation?. *The Quarterly Journal of Experimental Psychology Section A*, *39*(2), 211-251.

Glass, G.V., Peckham, P.D., & Sanders, J.R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237-288.

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). (Vol. 922). Hoboken, NJ: John Wiley & Sons.

Grim, P., Rosenberger, R., Rosenfeld, A., Anderson, B., & Eason, R. E. (2013). How simulations fail. *Synthese*, *190*(12), 2367-2390.

Hoedemaker, R. S., & Gordon, P. C. (2014). It takes time to prime: Semantic priming in the ocular lexical decision task. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(6), 2179-2197.

Hoedemaker, R. S., & Gordon, P. C. (2017). The onset and time course of semantic priming during rapid recognition of visual words. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(5), 881-902.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C. S., ... & Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods*, *45*(4), 1099-1114.

Judd, C.M., McClelland, G.H., & Culhane, S.E. (1995). Data analysis: Continuing issues in the everyday analysis of psychological data. *Annual Review of Psychology*, *46*, 433-465.

Kaplan, D. (1989). A study of the sampling variability and z-values of parameter estimates from misspecified structural equation models. *Multivariate Behavioral Research*, *24*, 41-57.

Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1457-1468.

Kinoshita, S., Forster, K. I., & Mozer, M. C. (2008). Unconscious cognition isn't that smart: Modulation of masked repetition priming effect in the word naming task. *Cognition*, *107*, 623–649.

Levine, D.W. & Dunlap W.P. (1982). Power of the F test with skewed data: Should one transform or not?. *Psychological Bulletin*, *92*(1), 272-280.

Lo, S., & Andrews, S. (2015). To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1-16.

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, *6*(3), 312-319.

Luce, R. D. (1986). *Response times* (No. 8). Oxford University Press.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203-208.

Maas, C. J., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*(2), 127-137.

Masson, M. E., & Kliegl, R. (2013). Modulation of additive and interactive effects in lexical decision by trial history. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 898-914.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/.

Raaijmakers, J. G., Schrijnemakers, J. M., & Gremmen, F. (1999). How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language*, *41*(3), 416-426.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59-108.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*, 510-532.

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, *83*(3), 190-214.

Ratcliff, R., Thapar, A., & McKoon, G. (2006). Aging and individual differences in rapid two-choice decisions. *Psychonomic Bulletin & Review*, *13*(4), 626-635.

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology*, *60*(3), 127-157.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 22nd ed.). Sage.

Schmider, E., Ziegler, M., Danay, E., Beyer, L., Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology: European Journal of Research Methods for the Behavioral & Social Sciences*, *6*(4), 147-151.

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H.-M., & Wittmann, W.W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *136*, 414-429.

Seltzer, M. H. (1993). Sensitivity analysis for fixed effects in the hierarchical model: A Gibbs sampling approach. *Journal of Educational and Behavioral Statistics, 18*(3), 207-235.

Sereno, S. C., Rayner, K., & Posner, M. I. (1998). Establishing a time-line of word recognition: evidence from eye movements and event-related potentials. *Neuroreport, 9*(10), 2195-2200.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and applied multilevel analysis* (Vol. 22nd ed.). Sage.

Speelman, C., & McGann, M. (2013). How mean is the mean?. *Frontiers in Psychology, 4*, 1-12. doi:10.3389/fpsyg.2013.00451.

Staub, A., White, S. J., Drieghe, D., Hollway, E. C., & Rayner, K. (2010). Distributional effects of word frequency on eye fixation durations. *Journal of Experimental Psychology: Human Perception and Performance, 36*(5), 1280-1293.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica, 30*, 276-315.

Tse, C.S., Balota, D.A., Yap, M.J. Duchek, J.M., & McCabe, D.P. (2010). Effects of healthy aging and early stage dementia of the Alzheimer's type on components of response time distributions in three attention tasks. *Neuropsychology, 24*, 300-315.

van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *The Quarterly Journal of Experimental Psychology, 47*(3), 631-650.

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). NY: Springer.

Wagenmakers, E. J., Krypotos, A. M., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: A survey of the field 33 years after Loftus. *Memory & Cognition, 40*(2), 145-160.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson tests of achievement.* Itasca, IL: Riverside Publishing.

Yap, M. J., & Balota, D. A. (2007). Additive and interactive effects on response time distributions in visual word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(2), 274-296.

Yap, M. J., Balota, D. A., & Tan, S. E. (2013). Additive and interactive effects in semantic priming: Isolating lexical and decision processes in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 140-158. doi:10.1037/a0028520

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance, 38*(1), 53-79.

Yap, M. J., Balota, D. A., Tse, C. S., & Besner, D. (2008). On the additive effects of stimulus quality and word frequency in lexical decision: evidence for opposing interactive influences revealed by RT distributional analyses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(3), 495-513.

Yap, M. J., Tse, C. S., & Balota, D. A. (2009). Individual differences in the joint effects of semantic priming and word frequency revealed by RT distributional analyses: The role of lexical integrity. *Journal of Memory and Language*, *61*(3), 303-325.